Noname manuscript No.
(will be inserted by the editor)

A penalty function approach for deriving optimality conditions in bilevel optimization

Rongzhu Ke · Christopher Thomas Ryan

Received: date / Accepted: date

Abstract We use a penalty function approach to derive necessary optimality conditions for nonconvex bilevel programs. The existing literature has concentrated on two basic approaches and their hybrids: (a) the KKT approach and (b) the value function approach that posits *calmness* conditions and employs nonsmooth analysis. Both methods reformulate the problem into a single level, possibly with complementarity or nonsmooth constraints. We explore an alternative approach based on a max-min reformulation. This produces parsimonious optimality conditions that involve a single alternate best response of the follower rather than an enumeration of best responses that is common in other methods. We provide examples where our optimality conditions hold but fail constraint qualifications of other approaches.

Key Words: Bilevel optimization, optimality conditions, penalty function

1 Introduction

s:introduction

We explore bilevel programming problems (BLPP) of the form

$$\begin{aligned} &\max_{x\in X,y} F(x,y) \\ \text{subject to}\;\; &y \in S(x) \qquad\qquad \text{(BLPP)} \\ &G(x,y) \geq 0 \end{aligned}$$

where X is a compact subset of $\mathbb{R}^n$ and $S(x)$ denotes the set of optimal solutions of the lower-level problem (LLP)

$$\begin{aligned} &\max_{y\in Y} f(x,y) \qquad\qquad \text{(LLP)} \\ &\text{s.t. } g(x,y) \geq 0 \end{aligned}$$

R. Ke
Hongkong Baptist University Department of Economics
E-mail: rongzhuke@hkbu.edu.hk

C.T. Ryan
University of Chicago Booth School of Business
E-mail: chris.ryan@chicagobooth.edu

with $F, f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, $G : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^q$, and $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^p$ where n, m, q and p are nonnegative integers and Y is a compact subset of $\mathbb{R}^m$.[1] All functions are continuously differentiable and f is twice continuously differentiable. There are no convexity requirements and, in particular, the lower-level problem need not be a convex optimization problem. We further assume that (BLPP) is feasible and possesses an optimal solution (x^*, y^*) that is interior to $X \times Y$. In fact, we will assume that $S(x^*)$ lies in the interior of Y.[2] Problems with equality constraints are easily adapted to our analysis, but we assume none are present for brevity. The over-arching goal of the paper is to provide optimality conditions that are satisfied by (x^*, y^*).

Bilevel programs are challenging. Their study inherently involves parametric, nonconvex and nonsmooth optimization at its core, due to the mathematical properties of the set $S(x)$ as a mapping of x. However, understanding bilevel optimization is worthwhile because of its numerous and far-reaching applications (for surveys of applications see [2,3]).

Considerable effort has been spent in deriving optimality conditions for (BLPP). Naïve application of standard single-level theory is well-documented to fail in general (see details in [5] and [15]). There are two broad approaches to deriving successful conditions. The first is classical and assumes that the lower-level problem is a convex optimization problem and hence replaceable by Karush-Kuhn-Tucker (KKT) conditions as constraints to the leader's problem (for a modern adaptation of the approach see, for instance, [4]). However, as demonstrated in [14], when the lower-level problem is not convex, the presence of nonoptimal stationary solutions to the lower-level problems implies that the resulting necessary optimality conditions do not characterize optimal solutions.

In response to these limitations, Ye and Zhu propose an alternate approach that applied to nonconvex bilevel programs in [15]. Their approach involves using the value function of the lower-level problem to create a single-level optimization problem. They give constraint qualifications for when Fritz-John (FJ) and KKT type necessary optimality conditions hold. One notable condition is *partial calmness*, which allows the value function to be handled in the objective of the resulting single-level problem rather than in the constraints. This yields clean optimality conditions that apply to a variety of cases. Later in [16], Ye and Zhu leverage both the KKT and value function approaches to yield new constraint qualifications and optimality conditions that apply even more broadly. Of note are *weak calmness* constraint qualifications based on the linearization cone of the bilevel problem with an additional constraint corresponding to the Lagrangian of the lower-level problem. Sufficient conditions to establish calmness are also provided. Since Ye and Zhu's two seminal papers, other classes of constraint qualifications have been discovered that build on both the KKT and value function approaches. Of note in this direction are papers by Dempe and co-authors (for instance, [4–6]).

Both the KKT approach and the value function approach convert (BLPP) to a single-level optimization problem. The resulting single-level optimization problems have additional complexity beyond a standard nonconvex optimization problem. In the KKT approach, complementarity constraints are included. In the value function approach, a nonsmooth function enters the constraints. Known results on complementarity and nonsmooth optimization are adapted to the bilevel setting to derive optimality conditions.

[1] In our formulation, both the leader and follower solve maximization problems. Much of the bilevel optimization literature has the leader and following solving minimization problems. Our preference for the maximization version comes from our grounding in application in contract theory, where the leader corresponds to a principal and the follower an agent, both of whom are utility maximizers. The direction of the constraints $G(x, y) \geq 0$ and $g(x, y) \geq 0$ (as opposed to $\leq$ constraints) is also typical of this setting, where $G(x, y) \geq 0$ corresponds to an individual rationality constraint that the agent must receive a utility of at least the utility of his best outside alternative.

[2] For clarity of exposition, we do not take up the issue of corner solutions in this paper. However, the theory presented here can be extended in a relatively straightforward manner to this setting by familiar techniques.

We analyze a new single-level reformulation of (BLPP). To do so, we draw inspiration from the classic paper of Mirrlees [14]. In this technique, the follower's optimization problem is replaced with infinitely-many *incentive compatibility* constraints of form $f(x, y) - f(x, z) \geq 0$ for all $z \in Y$ where $g(x, z) \geq 0$ to encode that (x, y) must be chosen with y as a best response to x. The derivation of an inner minimization in our max-min formulation derives from the fact that the optimization problem $\min_{z \in Y, g(x,z) \geq 0} f(x, y) - f(x, z)$ must have an optimal value greater than or equal to zero. We then employ an exact penalization method to the resulting max-min problem to derive necessary optimality conditions for the equivalent bilevel problem.

Penalty functions are an established tool to study bilevel problems. Most papers use them to design algorithms to solve bilevel programs numerically and focus on issues of exactness and convergence (see for example [7,12]). Liu et al. [10] derive optimality conditions for convex bilevel programs using penalty functions. Marcotte and Zhu [11] study generalized bilevel programs using penalty functions and derive optimality conditions under certain convexity conditions. By contrast, our focus is nonconvex problems. Moreover, Liu et al. [10] penalize the KKT reformulation where we penalize our max-min reformulation. We do not advocate the use of our penalty function for numerically solving bilevel programs. Indeed, our penalty function is tightly constructed to derive optimality conditions and hence takes on novel features compared to more standard penalty functions (these novel features are discussed at length in Section 2.1).

A point of differentiation between the optimality conditions that we derive and those of other authors is that our conditions involve a *single* alternate best response to the follower's optimal choice. Optimality conditions (3.13)-(3.16) in [16], by contrast, involve all alternate best responses. Much effort is made to focus our optimality condition on a single alternate best response. This is most clearly seen in the proofs of Theorem 1 and 2 and underlines several features of our penalty function (in particular, terms (vi)–(viii) seen in (6) below, whose delicate interplay is carefully studied in the proof of Theorem 1). Moreover, we study four examples in Section 3 where our optimality conditions apply but the constraint qualifications of other known optimality conditions fail, including those in [16].

The significance of developing an optimality condition with a single alternate best response can be seen in applications of these ideas to the study of principal-agent problems conducted by the authors. In [9], we show how a sparse optimality condition involving a Lagrange multiplier for a single alternate best response is critical in showing the monotonicity of optimal contracts. That paper uses a simpler version of the optimality condition studied here, where the follower's problem is unconstrained (that is, no g is present). In [8], the same sparse optimality condition is essential for providing a general method for solving moral hazard problems using such optimality conditions. It is not clear how conditions involving multiple best responses could be used to replicate these results.

Finally, optimality conditions based on calmness-like conditions typically have constraint qualifications with the following flavor: (i) they involve the leader's objective function and (ii) the qualification must be true for a collection of direction vectors. For instance, Theorem 3.2 in [16] requires the inner product of the leader and follower's gradients with a family of directions d from a cone to be nonnegative. By contrast, our constraint qualifications are more "classical" in form – they involve only the constraints of the leader's problem (this includes the follower's objective function) and the existence of a single direction d.

2 Deriving optimality conditions for bilevel programs

s:main-section

We begin by transforming the bilevel program to a "single-level" equivalent problem, albeit one with a "max-min" structure. This involves introducing an auxiliary decision variable z. First, observe that (BLPP) is equivalent to:

eq:IC-as-min eq:IR-again eq:lower-level-feasible

$$\begin{aligned} \max_{x,y} \quad & F(x,y) \\ \text{subject to} \quad & \min_{z:g(x,z)\geq 0} \{f(x,y) - f(x,z)\} \geq 0 && (1) \\ & G(x,y) \geq 0 && (2) \\ & g(x,y) \geq 0. && (3) \end{aligned}$$

The auxiliary variable z plays the role of an alternate choice for the follower. A first step is to pull the minimization operator out from the constraint (1) and behind the objective function. This requires handling the possibility that a choice of x does not implement y, in which case (1) is violated. To deal with this, we define the extended-real valued function from $X \times Y \times Y \to \mathbb{R} \cup \{-\infty\}$,

eq:FI

$$F^I(x,y,z) = \begin{cases} F(x,y) & \text{if } f^I(x,y) - f^I(x,z) \geq 0 \\ -\infty & \text{otherwise} \end{cases} \tag{4}$$

where

eq:fI

$$f^I(x,y) = \begin{cases} f(x,y) & \text{if } g(x,y) \geq 0 \\ -\infty & \text{otherwise} \end{cases} \tag{5}$$

for all $x \in X$ and $y, z \in Y$. The reformulation is:

$$\begin{aligned} \max_{x\in X, y\in Y} \; \min_{z\in Y} & F^I(x,y,z) \\ \text{subject to } & G(x,y) \geq 0 && \text{(Max-Min)} \\ & g(x,y) \geq 0. \end{aligned}$$

The objective $F^I(x,y,z)$ is well-defined (that is, we avoid $f^I(x,y) - f^I(x,z) = -\infty + \infty$) because of the constraint $g(x,y) \geq 0$.

lemma:equivalence-with-indicator

Lemma 1 *(Max-Min) and (BLPP) are equivalent problems in the sense that (a) if (x^*, y^*, z^*) is an optimal solution to (Max-Min) then (x^*, y^*) is an optimal solution to (BLPP), and (b) if (x^*, y^*) is an optimal solution to (BLPP) then (x^*, y^*, z) is an optimal solution to (Max-Min) for any $z \in Y$. In either case, (Max-Min) and (BLPP) have the same optimal objective value.*

Proof We use the following claim.

Claim 1 *Let $(\hat{x}, \hat{y})$ be such that $G(\hat{x}, \hat{y}) \geq 0$ and $g(\hat{x}, \hat{y}) \geq 0$. Then $(\hat{x}, \hat{y})$ is bilevel feasible if and only if $\min_z F^I(\hat{x}, \hat{y}, z) > -\infty$. Moreover, when $(\hat{x}, \hat{y})$ is bilevel feasible $\min_z F^I(\hat{x}, \hat{y}, z) = F(\hat{x}, \hat{y})$.*

We first prove the claim. Suppose $(\hat{x}, \hat{y})$ is not bilevel feasible. Then there exists a $\hat{z}$ with $g(\hat{x}, \hat{z}) \geq 0$ such that $f(\hat{x}, \hat{y}) < f(\hat{x}, \hat{z})$ and hence $F^I(\hat{x}, \hat{y}, \hat{z}) = -\infty$. This implies $\min_z F^I(\hat{x}, \hat{y}, z) = -\infty$.

Conversely, suppose $(\hat{x}, \hat{y})$ is bilevel feasible. For any z such that $g(\hat{x}, z) \ngeq 0$ we have $f^I(\hat{x}, z) = -\infty$ and so $f^I(x,y) - f^I(x,z) \geq 0$ is sure to hold, implying $F^I(\hat{x}, \hat{y}, z) = F(\hat{x}, \hat{y}) > -\infty$. On the other hand, if z satisfies $g(\hat{x}, z) \geq 0$ then $f^I(\hat{x}, \hat{y}) \geq f^I(\hat{x}, z)$ since $(\hat{x}, \hat{y})$ is bilevel feasible. This implies $F^I(\hat{x}, \hat{y}, z) = F(\hat{x}, \hat{y}) > -\infty$. This establishes the claim.

To establish part (a) of the lemma first note that val(Max-Min) $> -\infty$. This follows since a bilevel optimal solution $(\hat{x}, \hat{y})$ exists and so by the claim, val(Max-Min) $\geq \min_z F^I(\hat{x}, \hat{y}, z) = F(\hat{x}, \hat{y}) > -\infty$. Also by the claim, any choice in $\arg\max_{x,y} \min_z \left\{F^I(x,y,z) : G(x,y) \geq 0, g(x,y) \geq 0\right\}$

is bilevel feasible since all other choices have an objective value of $-\infty$. In particular, (x^*, y^*) is bilevel feasible. Since, again by the claim, $F^I(x,y,z) = F(x,y)$ for any (x,y) that are bilevel feasible, the optimality of (x^*, y^*, z^*) implies $F(x^*, y^*) = F^I(x^*, y^*, z^*) \geq F^I(x,y,z) = F(x,y)$ for any bilevel feasible (x,y). This means (x^*, y^*) is an optimal solution to (BLPP). The fact that val(Max-Min) = val(BLPP) then immediately follows from $F(x^*, y^*) = F^I(x^*, y^*, z^*)$.

Now for part (b). Since (x^*, y^*) is an optimal solution to (BLPP) then we know $G(x^*, y^*) \geq 0$, $g(x^*, y^*) \geq 0$ and (by the claim) $\min_z F^I(x^*, y^*, z) = F(x^*, y^*) > -\infty$. Hence any choice of $(x,y) \in \arg\max_{x,y} \min_z \left\{F^I(x,y,z) : G(x,y) \geq 0, g(x,y) \geq 0\right\}$ must be bilevel feasible and has (Max-Min) objective value $F(x,y)$. Since $F(x^*, y^*) \geq F(x,y)$ by bilevel optimality, this implies that

$$(x^*, y^*) \in \arg\max_{x,y} \min_z \left\{F^I(x,y,z) : G(x,y) \geq 0, g(x,y) \geq 0\right\}.$$

Note that for any choice of z, $F(x^*, y^*, z) = F(x^*, y^*)$. We conclude (x^*, y^*, z) is an optimal solution to (Max-Min) for any choice of z. □

2.1 A penalty function

ss:penalty-function

Let (x^*, y^*) be a bilevel optimal solution. Then by Lemma 1, (x^*, y^*, z^*) is an optimal solution to (Max-Min) for an arbitrary choice of $z^* \in S(x^*)$ not equal to y^*. Suppose also that the standard Mangasarian-Fromovitz Constraint Qualification (MFCQ) for the lower-level problem (given x^*) holds at z^*:

(Q1) There exists a $d_z \in \mathbb{R}^m$ such that $\nabla_y g_\ell(x^*, z^*)^\top d_z > 0$ for all $\ell \in A_g(x^*, z^*)$ where $A_g(x^*, z^*) = \{\ell : g_\ell(x^*, z^*) = 0\}$.[3]

We introduce the following penalty function:

$$\begin{aligned}
F^k(x,y,z) = F(x,y) &- \underbrace{\tfrac{k}{2}\sum_{j=1}^{q}(G_j^-(x,y))^2}_{\text{(i)}} - \underbrace{\tfrac{k}{4}\sum_{\ell=1}^{p}(g_\ell^-(x,y))^4}_{\text{(ii)}} - \underbrace{\tfrac{k}{2}\sum_{\ell \in A_g(x^*,z^*)}(g_\ell^+(x,z^*))^2}_{\text{(iii)}} \\
&- \underbrace{\tfrac{\alpha_1}{2}\|x - x^*\|^2}_{\text{(iv)}} - \underbrace{\tfrac{\alpha_2}{2}\|y - y^*\|^2}_{\text{(v)}} + \underbrace{\tfrac{k^{3/4}}{2}\|z - z^*\|^2}_{\text{(vi)}} - \underbrace{\tfrac{k}{4}||(x,y) - (x^*, y^*)||^2||z - z^*||^4}_{\text{(vii)}} \\
&- \underbrace{\tfrac{k}{4}\left(\min\left\{0, f(x,y) - \tfrac{k}{4}\sum_{\ell=1}^{p}(g_\ell^-(x,y))^4 - [f(x,z) - \tfrac{k}{4}\sum_{\ell=1}^{p}(g_\ell^-(x,z))^4 - \tfrac{\alpha_3}{2}\Delta(k)\|z - z^*\|^2]\right\}\right)^4}_{\text{(viii)}} \\
&- \underbrace{\tfrac{k}{2}(\min\{0, f(x,y) - \tfrac{k}{4}\sum_{\ell=1}^{p}(g_\ell^-(x,y))^4 - f(x,z^*)\})^2}_{\text{(ix)}} - \underbrace{\tfrac{k}{2}||\nabla_y f(x,y) - \nabla_y f(x^*, y^*)||^2}_{\text{(x)}}
\end{aligned} \quad (6)$$

eq:penalty-function

[3] There are several places in the course of the argument where (Q1) is critical. See, for instance, the proof of Claim 2 in Section A.1 and in the proof of Theorem 1, particularly in the proof of the needed Subclaim 5 in Section A.6. Assuming (Q1) holds allows us to leverage exactness properties of the penalty function $f^k(x^*, z) := f(x^*, z) - \frac{k}{4}\sum_{\ell=1}^{p}(g_\ell^-(x^*, z))^4 - \frac{\alpha}{2}\|z - z^*\|^2$ of the lower level problem (LLP). In particular, this gives us asymptotic properties of the penalty function $f^k(x^*, z)$ and ensures that $\max_z f^k(x^*, z) \to f(x^*, z^*)$. This, in turn, has implications for the asymptotic of the penalty function for the (Max-Min) problem defined in (6).

where α_1, α_2 and α_3 are positive constants, $g_\ell^-(x,z^*) = \min\{0, g_\ell(x,z^*)\}$, $g_\ell^+(x,z^*) = \max\{0, g_\ell(x,z^*)\}$, and $A_g(x^*,z^*) = \{\ell : g_\ell(x^*,z^*) = 0\}$. The value

eq:define-delta

$$\mathit{\Delta}(k) = \max_{z\in Y}[f(x^*,z) - \tfrac{k}{4}2^{-9}\sum_{\ell=1}^{p}(g_\ell^-(x^*,z))^{10}] - f(x^*,z^*). \tag{7}$$

Note that α_3 cannot be arbitrary and must be sufficiently large so that the MFCQ condition in (Q1) holds for solutions of the lower-level problem in a sufficiently small neighborhood of z^* (discussed in detail below). As we will see later, Corollary 2 below shows $\mathit{\Delta}(k) \to 0$ as $k \to \infty$.

Our approach is to take limits of the first-order conditions of $\max_{x,y}\min_z F^k(x,y,z)$ as $k \to \infty$. The key challenge is to then “evacuate” any conditions involving the auxiliary variable z to recover optimality conditions solely in x and y, the decision variables in (BLPP). Next, we introduce some additional notation and provide some outline of how our penalty function approach works. This discussion is at a high-level, with precise details to follow.

We define the function

eq:optimal-value-over-z

$$\varphi^k(x,y) := \min_{z\in Y} F^k(x,y,z) \tag{8}$$

and mapping

eq:optimal-set-over-z

$$\zeta^k(x,y) := \operatorname{argmin}_{z\in Y} F^k(x,y,z). \tag{9}$$

For all k we define iterates:

eq:define-x-k-y-k

$$(x^k,y^k) \in \operatorname{argmax}_{x\in X, y\in Y} \varphi^k(x,y) \text{ and } z^k \in \zeta^k(x^k,y^k). \tag{10}$$

Before analyzing φ^k and ζ^k further, we remark on a few of noticeable differences between this penalty function and a more standard penalty functions. First, since there is both a minimization and a maximization in (Max-Min), we have both positive and negative penalty terms. However, the simple difference in sign is not enough to disentangle these two optimizations problems and prioritize the minimization. These leads to additional complexity in the structure of the terms (especially terms (vi)-(viii)), which is most clearly seen in the proof of Theorem 1 below. A clear example of the sophistication of these penalty terms can be noticed in term (vi) having a different power of k than all other terms. This puts the priority first on the outer maximization in (Max-Min). This creates a structure where the inner minimization proceeds as a response to the outer maximization.

Before moving on, we provide some intuitive explanation of each of the penalty terms (i)-(x). Penalty function terms (i)–(ii) and (iv)–(v) are standard and closely mimic the single-level problem (as studied in Chapter 3 of [1]). The only difference in (ii) from the standard choice is that g_ℓ^- is raised to the fourth power. The choice is made to match the powers in terms (viii), where the fourth power is necessary to ensure twice continuous differentiability of F^k with respect to z. This property is critical in the proof of Lemmas 3 and 4.

Terms (vi)–(ix) are the key terms for encapsulating the (Max-Min) (and hence bilevel) structure. Term (vi) drives the sequence z^k towards z^*, allowing us to focus on a single alternate best response to x^*. Term (vii) is introduced to modulate the speed of convergence of x and y with respect the converge of z. It forces the z^k sequence to converge to z^* faster than (x^k,y^k) converges to (x^*,y^*), adhering to the fact that the minimization over z occurs first in (8). Term (viii) works to guarantee that lower-level optimality through penalizations of the lower-level problem. In particular, the $\mathit{\Delta}(k)$ plays a critical role in the proof of Claim 2 where its primary function is to drive the iterates z^k close to z^* in the lower level problem so that (Q1) can be used to control the asymptotic growth the g_ℓ^- terms.

Terms (iii) and (x) is present to help structure the constraint qualifications and optimality conditions that are eventually derived (see Theorems 4 and 5). It does not play a major role in the asymptotic reasoning that follows but shows up later in the constraint qualification and optimality condition developed in Section 2.4.

We must admit that this intuition is quite rough and the reader will appreciate that the penalty function (6) is highly tailored to meet the specific needs of deriving optimality conditions. We derived the precise form of the penalty function by a careful back-and-forth between the various needs of the arguments that follow. A precise understanding of the role of each penalty term can only really be appreciated by following the sequential steps of those arguments. To the extent possible, we point out where the structure of each penalty function term is especially relevant in the reasoning.

Returning to our notation φ^k, ζ^k, (x^k, y^k) and z^k defined in (8)–(10), it is not initially clear that φ^k is differentiable, a required property for us to “evacuate” z^k from our optimality conditions. We must understand how the optimal choice of z^k acts as a mapping of x and y. A key result in this regard is Lemma 4 below, which shows that, in fact, ζ^k is a *function* of (x, y) in a neighborhood sufficiently close to (x^k, y^k) when k is large.

Establishing this is one of the major technical challenges of the paper and can explain much of the complexity of the penalty function. The proof requires the careful construction of terms (vi)–(viii) in the penalty function that involve the choice z. Indeed, the difference in the powers of the k for these terms, it turns out, makes term (vi) dominate as k gets large ensuring the strict convexity asymptotically in z of term (vi)-(viii), which ensures a unique choice for z^k. The power of k in term (vi) can also not be too large in order to ensure that limit points of the sequence of (x^k, y^k) are bilevel feasible, a key part of the proof of Lemma 2 below.

If ζ^k were merely a set-valued mapping, it would make it difficult to derive optimality conditions for (x^k, y^k) that typically involve taking derivatives. The fact that ζ^k is a singleton allows us to write (x^k, y^k) as a local maximizer of $\varphi^k(x, y) = F^k(x, y, \zeta^k(x, y))$, where we have now handled the minimization operation that was complicating the definition of φ^k. The next key result (Theorem 2) is to show that φ^k is a *directionally differentiable* function on a sufficiently small neighborhood of (x^k, y^k), using this new expression for φ^k.

At this point, we can give a relatively straightforward optimality condition for (x^k, y^k) to the penalized problem: $\nabla_{x,y}\varphi^k(x^k, y^k) = 0$. The final remaining step is to observe that $\nabla_{x,y}\varphi^k(x^k, y^k) = \nabla_{x,y}F^k(x^k, y^k, \zeta^k(x^k, y^k))$ for (x^k, y^k) sufficiently close to (x^*, y^*). This is also achieved in Theorem 2, using an Envelope Theorem-type result. Finally, we argue that

eq:necessary-optimality-conditions

$$\lim_{k\to\infty} \nabla_{x,y}F^k(x^k, y^k, \zeta^k(x^k, y^k)) = 0 \tag{11}$$

provides necessary optimality conditions for (BLPP), as in standard penalty function methods for deriving optimality conditions. Equation (11) is the basis of our optimality conditions in Theorem 3–5 below.

2.2 Convergence

ss:convergence

Let $\hat{F}(k) := \max_{x,y}\min_z F^k(x, y, z) = F^k(x^k, y^k, z^k)$ denote the value of the penalized problem as a function of k and $F^* = F(x^*, y^*)$ denote the optimal value of (BLPP) and (Max-Min).

Lemma 2 (Exactness Lemma) lemma:exactness *Let (x^*, y^*, z^*) be a given optimal solution to (Max-Min) and z^* an element of $S(x^*)$ where (Q1) holds. Then, $\lim_{k\to\infty}\hat{F}(k) = F^*$.*

Proof We first show that $\liminf_{k\to\infty} \hat{F}(k) \geq F^*$ for all k. Letting $z^k_* \in \zeta^k(x^*, y^*)$ (and noting z^k_* need not equal z^* nor z^k) we have

$$\hat{F}(k) \geq \min_z F^k(x^*, y^*, z)$$

$$= F^k(x^*, y^*, z^k_*)$$

$$= F(x^*, y^*) + \tfrac{k^{3/4}}{2}\|z^k_* - z^*\|^2 - \tfrac{k}{4}\left(\min\left\{0, f(x^*, y^*) - [f(x^*, z^k_*) - \tfrac{k}{4}\sum_{\ell=1}^{p}(g^-_\ell(x^*, z^k_*))^4 - \tfrac{\alpha_3}{2}\Delta(k)\|z^k_* - z^*\|^2]\right\}\right)^4$$

where the last equality follows from the property that at (x^*, y^*) all constraints are satisfied (and so penalty terms (i)-(ii) disappear), $g_\ell(x^*, z^*) = 0$ for all ℓ in the sum in penalty function (iii) and so it disappears, and terms (vii), (ix) and (x) disappear when $(x, y) = (x^*, y^*)$. The only remaining penalty terms are (vi) and (viii). Observe that it suffices to show term (viii) goes to 0 as $k \to \infty$ since term (vi) is nonnegative. That is, we want to show

eq:last-thing-to-show

$$\tfrac{k}{4}\left(\min\left\{0, f(x^*, y^*) - [f(x^*, z^k_*) - \tfrac{k}{4}\sum_{\ell=1}^{p}(g^-_\ell(x^*, z^k_*))^4 - \tfrac{\alpha_3}{2}\Delta(k)\|z^k_* - z^*\|^2]\right\}\right)^4 \to 0. \tag{12}$$

Observe that

$$\left(\min\left\{0, f(x^*, y^*) - [f(x^*, z^k_*) - \tfrac{k}{4}\sum_{\ell=1}^{p}(g^-_\ell(x^*, z^k_*))^4 - \tfrac{\alpha_3}{2}\Delta(k)\|z^k_* - z^*\|^2]\right\}\right)^4$$

$$\leq \left(\max_z \left[f(x^*, z) - \tfrac{k}{4}\sum_{\ell=1}^{p}(g^-_\ell(x^*, z))^4 - \tfrac{\alpha_3}{2}\Delta(k)\|z - z^*\|^2 \right] - f(x^*, y^*) \right)^4 = (\rho(k) - f^*)^4$$

where

eq:rho

$$\rho(k) := \max_{z\in Y}\left[f(x^*, z) - \tfrac{k}{4}\sum_{\ell=1}^{p}(g^-_\ell(x^*, z))^4 - \tfrac{\alpha_3}{2}\Delta(k)\|z - z^*\|^2 \right] \tag{13}$$

and $f^* = f(x^*, y^*)$. The following claim shows how we can control the speed of this term, it is proven in Appendix A.1.

claim:speed-of-convergence

Claim 2 *If (Q1) holds then* $\rho(k) - f^* = O(k^{-1/3})$.

Thus, the left-hand side of (12) is $O(k(k^{-1/3})^4) = O(k^{-1/3})$ and (12) holds. In turn, $\liminf_{k\to\infty} \hat{F}(k) \geq F^*$ holds.

We now work to show a converse direction, namely that $\limsup_{k\to\infty} \hat{F}(k) \leq F^*$. We abuse notation so that k indexes a single convergent subsequence of the (x^k, y^k) with limit point (x_∞, y_∞). Under this convention, we will show $\lim_{k\to\infty} \hat{F}(k) \leq F^*$. If (x^k, y^k, z^k) is such that $F^k(x^k, y^k, z^k) \to -\infty$ then $\liminf_{k\to\infty} \hat{F}(k) \geq F^*$ for all k implies $F^* = -\infty$. However, we know (Max-Min) has a finite optimal value, so this is a contradiction. Suppose (x_∞, y_∞) is such that $G(x_\infty, y_\infty) \ngeq 0$. Then, by the continuity of G, we know $\lim_{k\to\infty} G(x^k, y^k) = G(x_\infty, y_\infty) \ngeq 0$. Hence, term (i) in (6) goes to $-\infty$ as $k \to \infty$. Only the positive term (vi) can reverse sending $F^k(x^k, y^k, z^k) \to -\infty$. However, term (vi) is $O(k^{3/4})$ since $||z^k - z^*||$ is uniformly bounded over the compact set Y, whereas term (i) is $\Omega(k)$ when $\lim_{k\to\infty} G(x^k, y^k) \not\to 0$. Hence, if $G(x_\infty, y_\infty) \ngeq 0$ then $F^k(x^k, y^k, z^k) \to -\infty$ as $k \to \infty$, a contradiction. We thus conclude $G(x_\infty, y_\infty) \geq 0$. Identical reasoning also implies $g(x_\infty, y_\infty) \geq 0$ and $g(x_\infty, z^*) \leq 0$, allowing us to drop terms (ii) and (iii).

Our focus next turns to penalty terms (vi)-(x). The following claim is proved in Appendix A.2.

claim:xfinty-works-with-ystar

Claim 3 (x_∞, y_∞) *is bilevel feasible. That is,* $g(x_\infty, y_\infty) \geq 0$ *and* $f(x_\infty, y_\infty) \geq f(x_\infty, z)$ *for all* z *with* $g(x_\infty, z) \geq 0$.[4]

[4] The proof of this claim essentially argues that term (vii) of the penalty function encodes bilevel feasibility, when used in combination with the other penalty function terms.

With the claim in hand, we can now establish $\lim_{k\to\infty} \hat{F}(k) \leq F^*$. Indeed,

eq:start

$$\lim_{k\to\infty} \hat{F}(k) \leq \lim_{k\to\infty} F^k(x^k, y^k, z^*) \tag{14}$$

$$\begin{aligned}
&= \lim_{k\to\infty} \Big(F(x^k, y^k) - \tfrac{\alpha_1}{2}\|x^k - x^*\|^2 - \tfrac{\alpha_2}{2}\|y^k - y^*\|^2 - \tfrac{k}{2}||\nabla_y f(x^k, y^k) - \nabla_y f(x^*, y^*)||^2 \\
&\quad - \tfrac{k}{4}(\min\{0, f(x^k, y^k) - \tfrac{k}{4}\sum_{\ell=1}^{p} g_\ell^-(x^k, y^k)^4 - [f(x^k, z^*) - \tfrac{k}{4}\sum_{\ell=1}^{p} g_\ell^-(x^k, z^*)^4]\})^4
\end{aligned}$$

eq:final-stuff

$$\tag{15}$$

$$- \tfrac{k}{4}(\min\{0, f(x^k, y^k) - \tfrac{k}{2}\sum_{\ell=1}^{p}(g_\ell^-(x^k, y^k))^4 - f(x^k, z^*)\})^2 \Big)$$

eq:second-direction

$$\leq F(x_\infty, y_\infty) \leq F(x^*, y^*) = F^* \tag{16}$$

where the first inequality follows from the definition of x^k and y^k and since z^* might not be in $\zeta^k(x_\infty, y_\infty)$. The equality comes from writing out $F^k(x^k, y^k, z^*)$ and dropping terms equal to 0 (this includes term (vii) which is zero when $z = z^*$). The second inequality arises from dropping negative terms and using the continuity of F to get $\lim_{k\to\infty} F(x^k, y^k) = F(x_\infty, y_\infty)$. The last inequality follows since (x^*, y^*) is an optimal solution to (BLPP) while (x_∞, y_∞) is a bilevel feasible solution by Claim 3. □

Using exactness, we can claim that all convergent subsequences of the iterates (x^k, y^k, z^k) converge to (x^*, y^*, z^*). The convergence of limits points of (x^k, y^k) immediately falls out of the proof of the Exactness Lemma.

In the remainder of the paper, we focus attention on a single convergent subsequence of the (x^k, y^k), which is guaranteed to exist by Bolzano-Weierstrass and the fact that this sequence lies in a compact set (although we may further refine that sequence on occasion). We abuse notation and continue to index that sequence by k. We denote the limit point of that sequence by (x_∞, y_∞).

theorem:convergence-main-variables

Corollary 1 *Let (x^*, y^*, z^*) be a given optimal solution to (Max-Min) and z^* in $S(x^*)$ where (Q1) holds. Then (x^k, y^k) converges to (x^*, y^*) as $k \to \infty$.*

Proof By exactness, (14)–(16) are equalities. Thus, all negative terms in (15) are zero. In particular, this means $||x_\infty - x^*|| = 0$ and $||y_\infty - y^*|| = 0$ and so $(x_\infty, y_\infty) = (x^*, y^*)$. □

Showing the convergence of the sequence z^k is a far more challenging task. Indeed, the convergence of z^k to z^* is the main challenge in deriving an optimality condition that involves the single alternate best response z^*. Consequently, the following result is the most challenging result to establish in this paper, despite its apparent simplicity. The challenge comes from the fact that we must directly deal with the positive penalty function term (vi) in combinations with the two negative terms (vii) and (viii) that involve z, which weakens the conclusions of exactness. The balancing act between these three terms (vi)–(viii) can explain much of their specific structure. The subtleties of these arguments are best seen in the proofs of Claims 4–6 in Appendices A.3–6 used in the proof below, where some of the more technical arguments in this paper reside.

theorem:z-converges-to-z-star

Theorem 1 *Let (x^*, y^*, z^*) be a given optimal solution to (Max-Min) and z^* in $S(x^*)$ where (Q1) holds. Let z^k be an element of $\zeta^k(x^k, y^k)$. Then every convergent subsequence of the z^k converges to z^*.[5] In particular, when restricting k to such subsequences, $||z^k - z^*|| = o(k^{-3/8})$.*

[5] By Bolzano-Weierstrass, z^k possesses at least one convergent subsequence.

Proof The basic starting fact that we have is that by the Exactness Lemma, we know that the sum of terms (vi)–(viii) of the penalty function $F^k(x^k, y^k, z^k)$ must converge to 0 as $k \to \infty$. We first show to leverage this property to show that every convergent subsequence of the z^k converges to z^*. Suppose otherwise, that there exists a subsequential limit $z_\infty \neq z^*$. Further abuse k to denote the index of that subsequence. Recall that Y is a compact set, so

eq:converge-to-constant

$$||z^k - z^*|| = \Theta(1) \tag{17}$$

since it cannot diverge. To build our contradiction, we start by examining term (viii) in some detail. To lighten notation we make the following definition:

eq:a-k-broken-up

$$A^k(z) = \min\{0, P^k + B^k(z)\} \tag{18}$$

where

$$P^k := f(x^k, y^k) - \tfrac{k}{4}\sum_{\ell=1}^{p}(g_\ell^-(x^k, y^k))^4 - f(x^k, z^*)$$

and

$$B^k(z) := f(x^k, z^*) - [f(x^k, z) - \tfrac{k}{4}\sum_{\ell=1}^{p}(g_\ell^-(x^k, z))^4 - \tfrac{\alpha_3}{2}\Delta(k)\|z - z^*\|^2].$$

This means term (viii) of the penalty function $F^k(x^k, y^k, z^k)$ is $-(k/4)A^k(z^k)^4$. The main work of the proof is to establish the following three claims, whose proofs are in Appendices A.3–6.

claim:z-infinity-in-interior

Claim 4 *$z_\infty \in S(x^*)$. In particular, for k sufficiently large, $z^k \in S(x^*) \subseteq \operatorname{int} Y$.*

Claim 4 allows us to work with first-order conditions in the lower-level problem to avoid challenges of corner solutions.

claim:term-viii-dominated-by-term-vi

Claim 5 *Term (viii) is dominated by term (vi), that is, term (viii) is $o(k^{3/4}||z^k - z^*||^2)$.*

With this claim in hand, the fact that terms (vi), (vii) and (viii) converge to 0, and term (vi) dominates term (viii) then the "race" is between terms (vi) and (vii). The next claim helps us understand the nature of that "race".

claim:internal-minimizer

Claim 6 *If Claim 5 holds – that is, term (viii) is $o(k^{3/4}||z^k - z^*||^2)$ – then*

eq:minimization-problem

$$\lim_{k\to\infty}\left(\tfrac{1}{2}||z^k - z^*||^2 - \tfrac{k^{1/4}}{4}||(x^k, y^k) - (x^*, y^*)||^2||z^k - z^*||^4\right)$$
$$= \lim_{k\to\infty}\min_{z\in Y}\left(\tfrac{1}{2}||z - z^*||^2 - \tfrac{k^{1/4}}{4}||(x^k, y^k) - (x^*, y^*)||^2||z - z^*||^4\right). \tag{19}$$

The argument of the minimization in (19) is the sum of terms (vi) and (vii) divided by $k^{3/4}$. The importance of this claim is that is tells that z^k is asymptotically optimal to a minimization over terms (vi) and (vii) and so the gradient (with respect to z) of those terms evaluated at z^k converges to 0 due to Claim 4.

Putting everything together, we know by exactness that the sum of terms (vi)-(viii) is

$$\tfrac{k^{3/4}}{2}\|z^k - z^*\|^2 - \tfrac{k}{4}\|d^k - d^*\|^2 \cdot \|z^k - z^*\|^4 - \tfrac{k}{4}(A^k(z^k))^4 \to 0,$$

where $d^k = (x^k, y^k)$ and $d^* = (x^*, y^*)$. Dividing this through by $k^{3/4}\|z^k - z^*\|^2$ (which is $\omega(1)$ so we are not dividing by zero) we have $\frac{1}{2} - \frac{k^{1/4}}{4}\|d^k - d^*\|^2 \cdot \|z^k - z^*\|^2 \to 0$ (term (viii) disappears by Claim 5). This implies

eq:to-contradict

$$k^{1/4}\|d^k - d^*\|^2\|z^k - z^*\|^2 \to 2. \tag{20}$$

However, by Claims 4 and 6, the first-order condition

eq:converge-the-other-way

$$(z_i^k - z_i^*)(1 - k^{1/4}\|d^k - d^*\|^2\|z^k - z^*\|^2) \to 0 \tag{21}$$

is satisfied in the limit, taking first-order conditions of the minimization problem in (19) with respect to z_i. If there exists at least one $z_i^k - z_i^* \nrightarrow 0$ then from (21) yields $1-k^{1/4}\|d^k - d^*\|^2\|z^k - z^*\|^2 \to 0$, contradicting (20). Thus, we must conclude the $z^k \to z^*$, as required.

It remains to show the "in particular" part of the proof that, in fact, $||z^k - z^*|| = o(k^{-3/8})$. This comes from showing that in fact term (vi) converges to zero in $F^k(x^k, y^k, z^k)$ and by rearranging gives $||z^k - z^*|| = o(k^{-3/8})$. This uses $||z^k - z^*|| \to 0$ that was established in the first part of the proof. Details of the "in particular" proof are placed in Appendix A.6. □

2.3 Differentiability

ss:differentiability

Having established convergence, we have hope to relate the optimality conditions of (x^*, y^*) to the optimality conditions enjoyed by (x^k, y^k) by passing through the limit. As described in Section 2.1, the challenge here is that these optimality conditions are dependent on the auxiliary variable z^k, which threatens differentiability properties of the associated functions when z^k is not unique given (x^k, y^k). The next results resolve these issues. We use the notation $B_r(m)$ to denote a ball of radius r centered at m, that is $B_r(m) := \{x : ||x - m|| < r\}$.

lemma:expand-singleton

Lemma 3 *For k sufficiently large, $\zeta^k(x^k, y^k)$ (as defined in (9)) is a singleton.*

Proof The proof relies on the following claim, proven in Appendix A.7. The idea of the proof of that claim is that among the penalty terms involving z^k, namely terms (vi)–(viii), term (vi) dominates for k sufficiently large and so

claim:second-derivative-diverges

Claim 7 *$F^k(x^k, y^k, z)$ is strictly convex in z on the set $B_{k^{-3/8}}(z^*)$ for k sufficiently large.*

Now, Theorem 1 says that $\zeta^k(x^k, y^k) \subseteq B_{\delta^k}(z^*)$ where δ^k is $o(k^{-3/8})$. Combined with Claim 7 this implies that for k sufficiently large, $\zeta^k(x^k, y^k)$ lies in a domain of strict convexity of F^k. A strict convex function has a unique minimizer, and so $\zeta^k(x^k, y^k)$ is a singleton for k sufficiently large. □

In fact, we show something even stronger. This result leverages the previous result and the continuity properties of the penalty function $F^k(x^k, y^k, z)$.

lemma:expand-singleton-more

Lemma 4 *Let (x^k, y^k) denote a sequence as defined in (10). For k sufficiently large, $\zeta^k(r^k, s^k)$ is a singleton for any sequence of choices (r^k, s^k) such that $(r^k, s^k) \in B_{k^{-2}}(x^k, y^k)$.*

Proof This result leverages the previous result and the continuity properties of the penalty function $F^k(x^k, y^k, z)$. Due the relative familiarity of this type of reasoning, we only provide a proof sketch. We first make the following claim.

claim:uniform-continuity

Claim 8 *For any sequence of choices (r^k, s^k) such that $(r^k, s^k) \in B_{k^{-2}}(x^k, y^k)$ we have*

eq:uniform-continuity

$$\sup_z |F^k(x^k, y^k, z) - F^k(r^k, s^k, z)| = o(1). \tag{22}$$

The challenge of Claim 8 is to show how to control the growth of each of the penalty terms as a function of k. This is complicated by the fact that most terms are multiplied by some power of k. Dealing with this growth in k explains why we assume $(r^k, s^k) \in B_{k^{-2}}(x^k, y^k)$ with radius k^{-2} and not radius k^{-1} (although this choice is somewhat arbitrary, any radius $o(k^{-1})$ suffices). The details of this argument mimic those found in the proof of Theorem 1. Further details are omitted.

With the claim in hand, we can leverage previous development to establish the result. By the Exactness Lemma and Claim 8 we have

eq:key-fact

$$\frac{k^{3/4}}{2}||\hat{z}^k - z^*||^2 - \frac{k}{4}\|(r^k, s^k) - (x^*, y^*)\|^2||\hat{z}^k - z^*||^4 - kA^k(\hat{z}^k|r^k, s^k)^4 \to 0 \tag{23}$$

for any $\hat{z}^k \in \zeta^k(r^k, s^k)$ where $A^k(\hat{z}^k|r^k, s^k)$ is the same as $A^k(z)$ but replacing (x^k, y^k) with (r^k, s^k). Property (23) is the key fact driving the proof of Theorem 1 and with (22) in hand we can show that any sequence $\hat{z}^k$ has $\hat{z}^k \to z^*$ and in particular $\frac{k^{3/4}}{2}||\hat{z}^k - z^*||^2 \to 0$. This, in turn, is the key fact driving the development of Lemma 3. Since the (r^k, s^k) converge sufficiently fast to the (x^k, y^k) this argument carries over, establishing that $\zeta^k(r^k, s^k)$ is a singleton.

theorem:varphi-differentiable

Theorem 2 *For k sufficiently large, $\varphi^k(x, y)$ is Gâteaux differentiable in x and y with for all $(\bar{x}, \bar{y}) \in B_{k^{-2}}(x^k, y^k)$ with Gâteaux derivative*[6] *$\nabla_{(x,y)}F^k(\bar{x}, \bar{y}, \zeta^k(\bar{x}, \bar{y}))$, where $\zeta^k(\bar{x}, \bar{y})$ is the unique optimal solution to $\min_z F^k(\bar{x}, \bar{y}, z)$.*

Proof Fix a k sufficiently large so that $\zeta^k(\bar{x}, \bar{y})$ is a singleton for every $(\bar{x}, \bar{y}) \in B_{k^{-2}}(x^k, y^k)$ (such a k is guaranteed by Lemma 4). Then ζ^k is a real-valued function (no longer set-valued) on the set $B_{k^{-2}}(x^k, y^k)$. Moreover, by the Theorem of Maximum, it is continuous on $B_{k^{-2}}(x^k, y^k)$.

Since $B_{k^{-2}}(x^k, y^k)$ is a full-dimensional open ball, given any direction $d = (d_x, d_y) \in X \times Y$ there exists an $\delta > 0$ such that $(\bar{x} + \epsilon d_x, \bar{y} + \epsilon d_y)$ remains in $B_{k^{-2}}(x^k, y^k)$ for any $0 < \epsilon < \delta$. Then for any such ϵ, $\zeta^k(\bar{x} + \epsilon d_x, \bar{y} + \epsilon d_y)$ is a real number and we can write:

$$\begin{aligned}
&\frac{F^k(\bar{x}+\epsilon d_x,\bar{y}+\epsilon d_y,\zeta^k(\bar{x}+\epsilon d_x,\bar{y}+\epsilon d_y))-F^k(\bar{x},\bar{y},\zeta^k(\bar{x}+\epsilon d_x,\bar{y}+\epsilon d_y))}{\epsilon}\\
&\le \frac{F^k(\bar{x}+\epsilon d_x,\bar{y}+\epsilon d_y,\zeta^k(\bar{x}+\epsilon d_x,\bar{y}+\epsilon d_y))-F^k(\bar{x},\bar{y},\zeta^k(\bar{x},\bar{y}))}{\epsilon}\\
&\le \frac{F^k(\bar{x}+\epsilon d_x,\bar{y}+\epsilon d_y,\zeta^k(\bar{x},\bar{y}))-F^k(\bar{x},\bar{y},\zeta^k(\bar{x},\bar{y}))}{\epsilon}
\end{aligned}$$

where both inequalities come from the definition of minimum. Sending $\epsilon \to 0$ from the positive direction through the above inequalities and using the continuity of ζ^k we find that

$$\lim_{\epsilon\to 0^+} \frac{F^k(\bar{x}+\epsilon d_x,\bar{y}+\epsilon d_y,\zeta^k(\bar{x}+\epsilon d_x,\bar{y}+\epsilon d_y))-F^k(\bar{x},\bar{y},\zeta^k(\bar{x}+\epsilon d_x,\bar{y}+\epsilon d_y))}{\epsilon} = \lim_{\epsilon\to 0^+} \frac{F^k(\bar{x}+\epsilon d_x,\bar{y}+\epsilon d_y,\zeta^k(\bar{x},\bar{y}))-F^k(\bar{x},\bar{y},\zeta^k(\bar{x},\bar{y}))}{\epsilon}.$$

Thus we can conclude $\lim_{\epsilon\to 0^+} \frac{\varphi^k(\bar{x}+\epsilon d_x,\bar{y}+\epsilon d_y)-\varphi^k(\bar{x},\bar{y})}{\epsilon} = \nabla_{(x,y)}F^k(\bar{x}, \bar{y}, \zeta^k(\bar{x}, \bar{y}))^\top h$.

A similar argument establishes that the left limit exists (taking $\epsilon \to 0^-$) and is also equal to $\nabla_{(x,y)}F^k(\bar{x}, \bar{y}, \zeta^k(\bar{x}, \bar{y}))^\top h$. This is true for any direction h and so φ^k is Gâteaux differentiable in x for all $(\bar{x}, \bar{y}) \in B_{k^{-2}}(x^k, y^k)$ with k sufficiently large. □

2.4 Optimality conditions

ss:optimality-conditions

Theorem 3 (Fritz-John-like condition) theorem:fritz-john-like-condition *Let (x^*, y^*, z^*) be a given optimal solution to (Max-Min) and z^* is in $S(x^*)$ where (Q1) holds. Then there exist nonnegative multipliers $\lambda_{G,j}$ for*

[6] It is straightforward to see that F^k is a continuously differentiable function and so $\nabla_{(x,y)}F^k$ exists.

$j = 1, \dots, q$, $\lambda_{g,\ell}$, κ_ℓ, and ρ_ℓ for $\ell = 1, \dots, p$, $\mu_0, \beta \geq 0$ and unsigned θ_i for $i = 1, \dots, m$, not all zero, such that:

$$0_n = \mu_0 \nabla_x F(x^*, y^*) + \sum_{j=1}^{q} \lambda_{G,j} \nabla_x G_j(x^*, y^*) + \sum_{\ell=1}^{p} (\lambda_{g,\ell} + \rho_\ell) \nabla_x g_\ell(x^*, y^*) \tag{24a}$$

$$- \sum_{\ell \in A_g(x^*,z^*)} \kappa_\ell \nabla_x g_\ell(x^*, z^*) + \sum_{i=1}^{m} \theta_i \nabla_x \frac{\partial f(x^*,y^*)}{\partial y_i} + \beta(\nabla_x f(x^*, y^*) - \nabla_x f(x^*, z^*))$$

$$0_m = \mu_0 \nabla_y F(x^*, y^*) + \sum_{j=1}^{q} \lambda_{G,j} \nabla_y G_j(x^*, y^*) + \sum_{\ell=1}^{p} (\lambda_{g,\ell} + \rho_\ell) \nabla_y g_\ell(x^*, y^*) \tag{24b}$$

$$+ \sum_{i=1}^{m} \theta_i \nabla_y \frac{\partial f(x^*,y^*)}{\partial y_i} + \beta \nabla_y f(x^*, y^*)$$

and the following complementary slackness conditions hold

$$\lambda_{G,j} G_j(x^*, y^*) = 0 \quad \text{for } j = 1, \dots, q, \tag{25a}$$

$$\lambda_{g,\ell} g_\ell(x^*, y^*) = 0 \quad \text{for } \ell = 1, \dots, p, \tag{25b}$$

$$\beta(f(x^*, y^*) - f(x^*, z^*)) = 0. \tag{25c}$$

Proof We focus on establishing the condition with respect to x (equation (24a)). The condition for y follows by analogous reasoning and is omitted for brevity.

In the previous two subsections, we established that (11) provides optimality conditions for (BLPP). Hence, we study the optimality conditions of the kth iterate of the penalized problem:

$$0 = \nabla_x \varphi^k(x^k, y^k) = \nabla_x F^k(x^k, y^k, z^k) \tag{26}$$

using Theorem 2. Expanding out (26) we get:

$$0 = \nabla_x F(x^k, y^k) - k \sum_{j=1}^{q} G_j^-(x^k, y^k) \nabla_x G_j(x^k, y^k) - k \sum_{\ell=1}^{p} g_\ell^-(x^k, y^k)^3 \nabla_x g_\ell(x^k, y^k)$$

$$- k \sum_{\ell \in A_g(x^*,z^*)} g_\ell^+(x^k, z^*) \nabla_x g_\ell(x^k, z^*) - \alpha_1(x^k - x^*) - \tfrac{k}{2} ||z^k - z^*||^4 (x^k - x^*)$$

$$- k\big(\min\big\{0, f(x^k, y^k) - \tfrac{k}{4} \sum_{\ell=1}^{p} (g_\ell^-(x^k, y^k))^4 - [f(x^k, z^k) - \tfrac{k}{4} \sum_{\ell=1}^{p} (g_\ell^-(x^k, z^k))^4 - \tfrac{\alpha_3}{2} \Delta(k) \|z^k - z^*\|^2]\big\}\big)^3 \times \tag{27a}$$

$$\left(\nabla_x f(x^k, y^k) - k \sum_{\ell=1}^{p} g_\ell^-(x^k, y^k)^3 \nabla_x g_\ell(x^k, y^k) - \nabla_x f(x^k, z^k) + k \sum_{\ell=1}^{p} g_\ell^-(x^k, z^k)^3 \nabla_x g_\ell(x^k, z^k) - \alpha_3 \Delta(k) ||z^k - z^*|| \right) \tag{27b}$$

$$- k \nabla_{yx}^2 f(x^k, y^k)(\nabla_y f(x^k, y^k) - \nabla_y f(x^*, y^*))$$

$$- k \left(\min\{0, f(x^k, y^k) - \tfrac{k}{4} \sum_{\ell=1}^{p} g_\ell^-(x^k, y^k)^4 - f(x^k, z^*)\} \right) \times$$

$$\left(\nabla_x f(x^k, y^k) - k \sum_{\ell=1}^{p} g_\ell^-(x^k, y^k)^3 \nabla_x g_\ell(x^k, y^k) - \nabla_x f(x^k, z^*) \right).$$

Observe that the term associated with penalty function term (vii) will disappear since, by Theorem 1, this term will converge to 0 as $k \to \infty$ since $||z^k - z^*|| = o(k^{-3/8})$. Moreover, we claim that the term in (27a)–(27b) also converges to 0 as $k \to \infty$ and hence can be removed from further development. Indeed, observe that the first component of the term (equation (27a) is precisely $kA^k(z^k)^3$ where A^k is defined in (18). Moreover, in Subclaim 4 in Appendix A.6, we show that A^k is $o(||z^k - z^*||)$. From Theorem 1 we know $||z^k - z^*||$ is $o(k^{-3/8})$. Putting these together implies $kA^k(z^k)^3$ is $o(k^{-1/8})$ and so converges to 0 as $k \to \infty$. Thus, we may re-express the above optimality condition as:

$$\begin{aligned}0 \leftarrow \nabla_x F(x^k,y^k) + \sum_{j=1}^{q} \eta_{G,j}^k \nabla_x G_j(x^k,y^k) + \sum_{\ell=1}^{p} \eta_{g,\ell}^k \nabla_x g_\ell(x^k,y^k) - \sum_{\ell \in A_g(x^*,z^*)} \xi_\ell^k \nabla_x g_\ell(x^k,z^*)\\ + \sum_{i=1}^{m} \vartheta_i^k \nabla_x \tfrac{\partial f(x^k,y^k)}{\partial y_i} + \gamma^k(\nabla_x f(x^k,y^k) - \nabla_x f(x^k,z^*)) + \sum_{\ell=1}^{p} \omega_\ell^k \nabla_x g_\ell(x^k,y^k) - \alpha_1(x^k - x^*)\end{aligned} \tag{28}$$

eq:writeout-nabla-g

as $k \to \infty$, where eq:define-coefficients

$$\eta_{G,j}^k := -kG_j^-(x^k,y^k) \qquad \text{for } j = 1,\dots,q, \tag{29a}$$

eq:G-sequence

$$\eta_{g,\ell}^k := -kg_\ell^-(x^k,y^k)^3 \qquad \text{for } \ell = 1,\dots,p, \tag{29b}$$

eq:g-y-sequence

$$\xi_\ell^k := kg_\ell^+(x^k,z^*) \qquad \text{for } \ell = 1,\dots,p, \tag{29c}$$

eq:g-z-sequence

$$\vartheta_i^k := -k(\tfrac{\partial}{\partial y_i} f(x^k,y^k) - \tfrac{\partial}{\partial y_i} f(x^*,y^*)) \qquad \text{for } i = 1,\dots,m, \tag{29d}$$

$$\gamma^k := -k(\min\{0, f(x^k,y^k) - \tfrac{k}{4}\sum_{\ell=1}^{p} g_\ell^-(x^k,y^k)^4 - f(x^k,z^*)\}) \text{ and} \tag{29e}$$

eq:f-g-sequence

$$\omega_\ell^k := \gamma^k \eta_{g,\ell}^k \qquad \text{for } \ell = 1,\dots,p. \tag{29f}$$

eq:product-sequence

Note the relationship between γ^k and term (ix) of the penalty function in (6). Denote

$$\delta^k = \sqrt{1 + \sum_{j=1}^{q}(\eta_{G,j}^k)^2 + \sum_{\ell=1}^{p}(\eta_{g,\ell}^k)^2 + \sum_{\ell=1}^{p}(\xi_\ell^k)^2 + \sum_{i=1}^{m}(\vartheta_i^k)^2 + (\gamma^k)^2 + \sum_{\ell=1}^{p}(\omega_\ell^k)^2} \tag{30}$$

eq:delta-sequence

and set $\mu_0^k := \frac{1}{\delta^k}$, $\lambda_{G,j}^k := \frac{\eta_{G,j}^k}{\delta^k}$ for $j = 1,\dots,q$, $\lambda_{g,\ell}^k := \frac{\eta_{g,\ell}^k}{\delta^k}$, $\kappa_\ell^k := \frac{\xi_\ell^k}{\delta^k}$, and $\rho_\ell^k := \frac{\omega_\ell^k}{\delta^k}$ for $\ell = 1,\dots,p$, $\theta_i^k = \frac{\vartheta_i^k}{\delta^k}$ for $i = 1,\dots,m$, and $\beta^k = \frac{\gamma^k}{\delta^k}$. Dividing (28) by δ^k yields

$$\begin{aligned}0 \leftarrow \mu_0^k \nabla_x F(x^k,y^k) + \sum_{j=1}^{q} \lambda_{G,j}^k \nabla_x G_j(x^k,y^k) + \sum_{\ell=1}^{p} \lambda_{g,\ell}^k \nabla_x g_\ell(x^k,y^k) - \sum_{\ell \in A_g(x^*,z^*)} \kappa_\ell^k \nabla_x g_\ell(x^k,z^*)\\ + \sum_{i=1}^{m} \theta_i^k \nabla_x \tfrac{\partial f(x^k,y^k)}{\partial y_i} + \beta^k(\nabla_x f(x^k,y^k) - \nabla_x f(x^k,z^*)) + \sum_{\ell=1}^{p} \rho_\ell^k \nabla_x g_\ell(x^k,y^k) - \tfrac{\alpha_1}{\delta^k}(x^k - x^*)\end{aligned} \tag{31}$$

eq:almost-all-the-way

as $k \to \infty$. From (30) we have

$$(\mu_0^k)^2 + \sum_{j=1}^{q}(\lambda_{G,j}^k)^2 + \sum_{\ell=1}^{p}(\lambda_{g,\ell}^k)^2 + \sum_{\ell=1}^{p}(\kappa_\ell^k)^2 + \sum_{i=1}^{m}(\theta_i^k)^2 + (\beta^k)^2 + \sum_{\ell=1}^{p}(\rho_\ell^k)^2 = 1, \tag{32}$$

eq:sum-to-one

which implies that the sequence $(\mu_0^k, \lambda_G^k, \lambda_g^k, \kappa^k, \theta^k, \beta^k, \rho^k)_{k=1}^{\infty}$ is bounded and so (since a sequence in Euclidean space) contains a convergent subsequence to some limit $(\mu_0, \lambda_G, \lambda_g, \kappa, \theta, \beta, \rho)$. From (32) we know the limit vector $(\mu_0, \lambda_G, \lambda_g, \kappa, \theta, \beta, \rho)$ is not the zero vector.

Now, redefine the sequence k to index this convergent subsequence and take $k \to \infty$ on both sides of (31). The continuous differentiability of F, G, f and g and the fact that $(x^k, y^k) \to (x^*, y^*)$ (from Corollary 1) yields (24a) (after some rearranging) and noting the fact $\frac{\alpha_1}{\delta^k}(x^k - x^*) \to 0$. The nonnegativity of $(\mu_0, \lambda_G, \lambda_g, \kappa, \beta, \rho)$ follows from (29a)–(29f). The complementarity conditions follow by standard arguments that we will not repeat here (see Proposition 3.3.5 in Bertsekas [1]). □

rem:lower-level-CQ-needed-for-FJ

Remark 1 Observe that constraint qualification (Q1) for the lower-level problem is needed for the Fritz-John-like condition. This is not the case for single-level optimization, where constraint qualifications are typically only needed to ensure the Fritz-John coefficient μ_0 is not zero.

To ensure that the Fritz-John coefficient μ_0 is nonzero, we introduce the following constraint qualifications. The first set of conditions assumes there exists a $z^* \in S(x^*)$ such that (Q1) holds and there exists a $d = (d_x, d_y) \in \mathbb{R}^n \times \mathbb{R}^m$ such that:

(Q2) $\nabla G_j(x^*, y^*)^\top d > 0$ for all $j \in A_G(x^*, y^*)$,
(Q3) $\nabla g_\ell(x^*, y^*)^\top d > 0$ for all $\ell \in A_g(x^*, y^*)$,
(Q4) $\nabla_x g_\ell(x^*, z^*)^\top d_x < 0$ for all $\ell \in A_g(x^*, z^*)$,
(Q5) $\nabla \frac{\partial f(x^*,y^*)}{\partial y_i}^\top d = 0$ for all $i = 1, \dots, m$,
(Q6) $\nabla f(x^*, y^*)^\top d \geq \nabla f(x^*, z^*)^\top d$ and $\nabla f(x^*, y^*)^\top d > \nabla f(x^*, z^*)^\top d$ if $A_G(x^*, y^*)$, $A_g(x^*, y^*)$ and $A_g(x^*, z^*)$ are empty, and
(Q7) $\nabla_y f(x^*, y^*)^\top d_y \geq 0$,

where $A_G(x^*, y^*) = \{j : G_j(x^*, y^*) = 0\}$, $A_g(x^*, y^*) = \{\ell : g_\ell(x^*, y^*) = 0\}$, and $A_g(x^*, z^*) = \{\ell : g_\ell(x^*, z^*) = 0\}$. Moreover,

(Q8) the columns of $\nabla^2_{yy} f(x^*, y^*))$ are linearly independent, and
(Q9) $||\nabla_x f(x^*, y^*) - \nabla_x f(x^*, z^*)|| \neq 0$.

theorem:mcfq-type-result-first

Theorem 4 *Let (x^*, y^*) be an optimal solution to (BLPP) and $z^* \in S(x^*)$. Suppose (Q1)–(Q9) hold. Then, there exist nonnegative multipliers $\lambda_{G,j}$ for $j = 1, \dots, q$, $\lambda_{g,\ell}$, ρ_ℓ and κ_ℓ for $\ell = 1, \dots, p$, $\beta \geq 0$ and unsigned θ_i for $i = 1, \dots, m$ such that:* eq:optimality-conditions

eq:optimality-conditions-x

$$\begin{aligned} 0_n = {} & \nabla_x F(x^*, y^*) + \sum_{j=1}^{q} \lambda_{G,j} \nabla_x G_j(x^*, y^*) + \sum_{\ell=1}^{p} (\lambda_{g,\ell} + \rho_\ell) \nabla_x g_\ell(x^*, y^*) \\ & - \sum_{\ell \in A_g(x^*, z^*)} \kappa_\ell \nabla_x g_\ell(x^*, z^*) + \sum_{i=1}^{m} \theta_i \nabla_x \frac{\partial f(x^*,y^*)}{\partial y_i} + \beta(\nabla_x f(x^*, y^*) - \nabla_x f(x^*, z^*)) \end{aligned} \tag{33a}$$

eq:optimality-conditions-y

$$\begin{aligned} 0_m = {} & \nabla_y F(x^*, y^*) + \sum_{j=1}^{q} \lambda_{G,j} \nabla_y G_j(x^*, y^*) + \sum_{\ell=1}^{p} (\lambda_{g,\ell} + \rho_\ell) \nabla_y g_\ell(x^*, y^*) \\ & + \sum_{i=1}^{m} \theta_i \nabla_y \frac{\partial f(x^*,y^*)}{\partial y_i} + \beta \nabla_y f(x^*, y^*) \end{aligned} \tag{33b}$$

and the complementary slackness conditions (25) *hold.*

Proof It suffices to show $\mu_0 \neq 0$ (and hence $\mu_0 > 0$ since $\mu_0 \geq 0$) in (24a) and (24b). Indeed, if $\mu_0 \neq 0$ then one can divide (24a) and (24b) through by μ_0 to yield (33a)–(33b). Suppose, by way of contradiction, that $\mu_0 = 0$. We first claim that at least one of λ_{G_j}, $\lambda_{g,\ell}$, ρ_ℓ, κ_ℓ, and β are positive (recall all these coefficients are nonnegative). Otherwise, we have $0_m =$

$\sum_{i=1}^{m} \theta_i \nabla_y \frac{\partial f(x^*,y^*)}{\partial y_i}$, which violates (Q8). We next claim that $\lambda_{G,j}, \lambda_{g,\ell}$, ρ_ℓ, and κ_ℓ are all zero. If not, then multiplying (24a) and (24b) through by the d given in (Q2)-(Q7) yields

$$
\begin{aligned}
0 = &\sum_{j=1}^{q} \lambda_{G,j} \nabla G_j(x^*,y^*)^\top d + \sum_{\ell=1}^{p} (\lambda_{g,\ell} + \rho_\ell) \nabla g_\ell(x^*,y^*)^\top d \\
&- \sum_{\ell \in A_g(x^*,z^*)} \kappa_\ell \nabla_x g_\ell(x^*,z^*)^\top d_x + \sum_{i=1}^{m} \theta_i \nabla \frac{\partial f(x^*,y^*)}{\partial y_i}^\top d \\
&+ \beta(\nabla_x f(x^*,y^*) - \nabla_x f(x^*,z^*))^\top d_x + \beta \nabla_y f(x^*,y^*)^\top d_y.
\end{aligned}
$$

Then from (Q2)–(Q7) and the fact that not all of $\lambda_{G,j}, \lambda_{g,\ell}$, ρ_ℓ, and κ_ℓ are zero, gives a contradiction of $0 > 0$ in the above expression. More directly, it shows $\lambda_{g,\ell} + \rho_\ell = 0$ and this, in turn, implies $\lambda_{g,\ell}$ and ρ_ℓ are both zero since these coefficients are nonnegative. Note that if (Q6) holds strictly we may also conclude that $\beta = 0$. However, we have already eliminated this possibility (since at least one of λ_{G_j}, $\lambda_{g,\ell}$, κ_ℓ, ρ_ℓ and β must be positive). Hence, to avoid contradiction we assume (Q6) holds with equality and $\beta > 0$.

In fact, we show something stronger, that $\eta^k_{G,j}$, $\eta^k_{g,\ell}$, ω^k_ℓ, and ξ^k_ℓ (as defined in (29)) are all bounded. Indeed, suppose otherwise, then dividing (28) by the norm of $(\eta^k_G, \eta^k_g, \omega^k, \xi^k)$ we get a similar contradiction as above when multiplying through by d. There are two remaining cases to consider.

Case 1: $\beta > 0$ and $\theta_i \neq 0$ for some i.

Since $\beta > 0$ and we must have $\gamma^k \to \infty$, where γ^k is as defined in (29e). Indeed, γ^k weakly dominates all other terms in the definition of δ^k in (30). Now, return to the first-order condition $\nabla_y F^k(x^k, y^k, z^k) = 0$ which is parallel to (28) but for y:

$$
\begin{aligned}
&\nabla_y F(x^k,y^k) + \sum_{j=1}^{q} \eta^k_{G,j} \nabla_y G_j(x^k,y^k) + \sum_{\ell=1}^{p} \eta^k_{g,\ell} \nabla_x g_\ell(x^k,y^k) \\
&+ \sum_{i=1}^{m} \vartheta^k_i \nabla_y \frac{\partial f(x^k,y^k)}{\partial y_i} + \gamma^k \nabla_y f(x^k,y^k) + \sum_{\ell=1}^{p} \omega^k_\ell \nabla_y g_\ell(x^k,y^k) \to 0
\end{aligned}
\qquad \text{eq:foc-y} \tag{34}
$$

as $k \to \infty$. Since $\omega_\ell = 0$ and $\gamma^k \to \infty$ we must have $\eta^k_{g,\ell} \to 0$ and so continuing from (34) yields

$$
\nabla_y F(x^k,y^k) + \sum_{j=1}^{q} \eta^k_{G,j} \nabla_y G_j(x^k,y^k) + \sum_{i=1}^{m} \vartheta^k_i \nabla_y \frac{\partial f(x^*,y^*)}{\partial y_i} + \gamma^k \nabla_y f(x^k,y^k) \to 0 \qquad \text{eq:foc-y-with-work-done} \tag{35}
$$

as $k \to \infty$. The next step is to argue that $\gamma^k \nabla_y f(x^k,y^k) \to 0$. This follows immediately by exactness, specifically the fact that term (ix) of the penalty function converges to 0. It remains to argue that $\nabla_y f(x^k,y^k)$ converges to 0 sufficiently fast. We will show this by leveraging the Exactness Lemma and term (xii) of the penalty function.

Note that if $\nabla_y f(x^*,y^*) = 0$ then exactness implies (from term (x)) that $\frac{k}{2}||\nabla_y f(x^k,y^k)||^2 \to 0$ and so $\nabla_y f(x^k,y^k)$ converges to 0 in order at least $k^{-1/2}$. This guarantees that $\gamma^k \nabla_y f(x^k,y^k) \to 0$. It remains to argue that $\nabla_y f(x^*,y^*) = 0$. Suppose not, then for k sufficiently large we have $\frac{\partial f(x^*,y^*)}{\partial y_i} \neq 0$ for at least one i. Fix one such dimension i. Define an alternate sequence $\tilde{z}^k$ where $\tilde{z}^k_i = y^k_i + \epsilon^k$ and $\tilde{z}^k_{-i} = y^k_{-i}$ and let $\epsilon^k = \Theta(k^{-1/20})$. We now claim that $F^k(x^k, y^k, \tilde{z}^k) \to -\infty$, a contradiction of the Exactness Lemma. Indeed, consider term (viii) of $F^k(x^k, y^k, \tilde{z}^k)$. The inside of term (viii), noting by Corollary 2 in Appendix A.1 that $\Delta(k) \to 0$, is:

eq:continue-from-here

$$\begin{aligned} & f(x^k, \tilde{z}^k) - \tfrac{k}{4} \sum_{\ell=1}^{p} g_\ell^-(x^k, \tilde{z}^k)^4 - [f(x^k, y^k) - \tfrac{k}{4} \sum_{\ell=1}^{p} g_\ell^-(x^k, y^k)^4] \\ &= \tfrac{\partial f(x^k y^k)}{\partial y_i} \epsilon^k - k \sum_{\ell=1}^{p} g_\ell^-(x^k, y^k)^3 \tfrac{\partial g_\ell(x^k y^k)}{\partial y_i} \epsilon^k + o(\epsilon^k) \end{aligned} \tag{36}$$

where we take the Taylor expansion around y^k at $\tilde{z}^k$. Recall that $\eta_{g,\ell}^k = -k g_\ell^-(x^k, y^k)^3 \to 0$ in (36) and thus continuing from (36) we have

$$f(x^k, \tilde{z}^k) - \tfrac{k}{4} \sum_{\ell=1}^{p} g_\ell^-(x^k, \tilde{z}^k)^4 - [f(x^k, y^k) - \tfrac{k}{4} \sum_{\ell=1}^{p} g_\ell^-(x^k, y^k)^4] = \tfrac{\partial f(x^k, y^k)}{\partial y_i} \epsilon^k + o(\epsilon^k) = \Theta(k^{-1/20}).$$

This implies that term (viii) in the penalty function is $\Theta(k \cdot (k^{-1/20})^4) = \Theta(k^{4/5})$. This dominates term (vi) of the penalty function, which has order $k^{3/4} ||\tilde{z}^k - z^*||^2 = O(k^{3/4})$ since $||\tilde{z}^k - z^*||$ is uniformly bounded above for all k since Y is a compact set. This implies that $F^k(x^k, y^k, \tilde{z}^k) \to -\infty$, contradicting the Exactness Lemma. Hence, returning to (35) we can conclude that the second term of the expression converges to 0 on its own as $k \to \infty$ and so

$$\nabla_y F(x^k, y^k) + \sum_{j=1}^{q} \eta_{G,j}^k \nabla_y G_j(x^k, y^k) + \sum_{i=1}^{m} \vartheta_i^k \nabla_y \tfrac{\partial f(x^*, y^*)}{\partial y_i} \to 0.$$

Dividing the above by $||\vartheta^k||$ yields (recall $\eta_{G,j}^k$ is a bounded sequence) and sending $k \to \infty$ yields:

$$\sum_{i=1}^{m} \tfrac{\vartheta_i^k}{||\vartheta^k||} \nabla_y \tfrac{\partial f(x^k, y^k)}{\partial y_i} = 0,$$

violating constraint qualification (Q8). Hence, Case 1 leads to a contradiction.

Case 2: $\beta > 0$ and $\theta_i = 0$ for all i.

In this case, $\gamma^k \to \infty$ while all other coefficients in (30) are bounded. This implies that term (ix) of the penalty function is a dominant term and hence by a perturbation argument similar to that of the proof of Theorem 1, we may conclude from term (ix) that $\frac{\partial f(x^k, y^k)}{\partial x_i} + \sum_{\ell=1}^{p} \eta_{g,\ell}^k \frac{\partial g_\ell(x^k, y^k)}{\partial x_i} - \frac{\partial f(x^k, z^*)}{\partial x_i} \to 0$ and $\frac{\partial f(x^k, y^k)}{\partial y_i} + \sum_{\ell=1}^{p} \eta_{g,\ell}^k \frac{\partial g_\ell(x^k, y^k)}{\partial y_i} \to 0$. As in the previous case since $\gamma^k \to \infty$ we know $\eta_{g,\ell}^k \to 0$ and hence for all i: $\frac{\partial f(x^k, y^k)}{\partial x_i} - \frac{\partial f(x^k, z^*)}{\partial x_i} \to 0$. This is ruled out by (Q9). □

We now state alternate constraint qualifications and optimality conditions for when (Q9) fails to hold.

theorem:mcfq-type-result-second

Theorem 5 *Let (x^*, y^*) be an optimal solution to (BLPP) and $z^* \in S(x^*)$. Suppose (Q1)–(Q8) hold but (Q9) does not hold. Then the negation of (Q9) and* (33b) *are necessary optimal conditions for (x^*, y^*). Moreover, if there does not exist a nonzero vector $\hat{d}$ (with $||\hat{d}|| = 1$) such* *that* eq:normalization-property

eq:cannot-d-G
$$\hat{d}^\top \nabla^2 G_j(x^*,y^*)^\top \hat{d} = \nabla G_j(x^*,y^*)^\top \hat{d} = 0 \text{ for every } j \in A_G(x^*,y^*) \tag{37a}$$

eq:cannot-d-g-x
$$\hat{d}_x^\top \nabla^2_{xx} g_\ell(x^*,z^*)^\top \hat{d}_x = \nabla_x g_\ell(x^*,z^*)^\top \hat{d}_x = 0 \text{ for every } \ell \in A_g(x^*,z^*) \tag{37b}$$

eq:cannot-d-f
$$\left(\nabla \frac{\partial f(x^*,y^*)}{\partial y_i}\right)^\top \hat{d} = 0 \text{ for every } i = 1,...,m, \text{ and} \tag{37c}$$

eq:cannot-d-g-all
$$\nabla g_\ell(x^*,y^*)^\top \hat{d} = 0 \text{ for every } \ell \in A_g(x^*,y^*), \tag{37d}$$

then there exist nonnegative multipliers $\lambda_{G,j}$ for $j = 1,\dots,q$, $\lambda_{g,\ell}$, and κ_ℓ for $\ell = 1,\dots,p$, $\beta \geq 0$ and unsigned θ_i for $i = 1,\dots,m$ such that optimality conditions (33) *hold, as well as complementary slackness conditions* (25) *hold.*

Proof In the proof of Theorem 4, note that (Q1)–(Q8) is sufficient to establish the optimality condition for y given by (33b). To establish the "moreover", we assume, by way of contradiction, that there exist no such optimality conditions of the form (33a)–(33b) with $\mu_0 \neq 0$. In particular, we may assume that the μ_0 as defined in Theorem 3 is zero.

The argument here is identical to that of Theorem 4 up until the end of Case 1, as this argument only depended on (Q1)-(Q8). Thus, we are in the situation of case 2: $\beta > 0$ and $\theta_i = 0$ for all i. This implies that $\gamma^k \to \infty$ and the other multipliers $\eta^k_{G,j}$, $\eta^k_{g,\ell}$, κ^k_ℓ and ω^k_ℓ are bounded sequences. This further implies that $\eta^k_{g,\ell} \to 0$.

To handle Case 2 when (Q9) fails, we require three technical claims, the proofs of which are found in the Appendix A.8, A.9, and A.10. These results refer to the sequence of vectors $u^k = (x^k - x^*, y^k - y^*)$.

claim:control-term-ix

Claim 9 $\gamma^k||\nabla_x f(x^k,y^k) - \nabla_x f(x^k,z^*) + \sum_{\ell=1}^p \eta^k_{g,\ell}\nabla_x g_\ell(x^k,y^k)|| = \Theta(1)$.

claim:control-gamma

Claim 10 $\gamma^k = O(k^{1/2}||u^k||^{1/2})$.

claim:d-k-gets-large

Claim 11 $k||u^k||^3 = \Omega(1)$.

With Claims 9–11 in hand, we go back to the two first-order conditions with respect to x and y in (27a)–(27b) and (34). For brevity, we only treat the conditions with respect to x. From Claim 10 we have $\gamma^k \to \infty$ and $k||u^k|| = \Omega((\gamma^k)^2)$ and thus $k||u^k|| \to \infty$. So, dividing the first-order condition $\nabla_x F^k(x^k,y^k,z^k) = 0$ in (28) by $k\|u^k\|$ yields, as $k \to \infty$,

eq:above-condition
$$0 \leftarrow \sum_{j=1}^q \frac{\eta^k_{G,j}}{k\|u^k\|}\nabla_x G_j(x^k,y^k) + \sum_{\ell=1}^p \frac{\eta^k_{g,\ell}}{k\|u^k\|}\nabla_x g_\ell(x^k,y^k) - \sum_{\ell\in A_g(x^*,z^*)} \frac{\xi^k_\ell}{k\|u^k\|}\nabla_x g_\ell(x^k,z^*) + \sum_{i=1}^m \frac{\vartheta^k_i}{k\|u^k\|}\nabla_x \frac{\partial f(x^k,y^k)}{\partial y_i} \tag{38}$$

using Claim 10 to see that $\frac{\gamma^k}{k\|u^k\|} = O(k^{-1/2}||u^k||^{-1/2}) = o(1)$ since we have assumed $\gamma^k \to \infty$ and so certainly $k^{1/2}||u^k||^{1/2} \to \infty$. Observe that all coefficients on the gradients in (38) must converge to zero to avoid contradiction. Otherwise, if we multiply through by the d furnished (Q1)–(Q7), we get (in the limit) $0 > 0$, a contradiction. This allows us to gain further insight. Consider, for example, the coefficient $\eta^k_{G,j}$ for $j \in A_G(x^*,y^*)$ (observe that other j's play no role). Since $\eta^k_{G,j}$ converges to 0 as k goes to infinity, we have that $\eta^k_{G,j} = -kG_j(x^k,y^k)$ for sufficiently large k (recall that $\eta^k_{G,j}$ is defined in (29a)). By taking the Taylor series expansion of G_j around (x^*,y^*), we have

$$\frac{\eta^k_{G,j}}{k\|u^k\|} = \frac{\nabla G_j(x^*,y^*)^\top u^k + o(u^k)}{\|u^k\|} \to \nabla G_j(x^*,y^*)^\top \hat{d} = 0,$$

where we let $\hat{d} = \lim_{k\to\infty} \frac{u^k}{\|u^k\|}$ and using the fact that $G_j(x^*, y^*) = 0$ since $j \in A_G(x^*, y^*)$. Similarly, we have $\nabla_x g_\ell(x^*, z^*)^\top \hat{d}_x = 0$ for every $\ell \in A_g(x^*, z^*)$ and $\left(\nabla \frac{\partial f(x^*, y^*)}{\partial y_i}\right)^\top \hat{d} = 0$ for every $i = 1, ..., m$. This establishes part of the contradiction, to conditions (37b) and (37c). By a similar line of reasoning, we can further divide the first-order condition $\nabla F^k(x^k, y^k, z^k)$ by $k\|u^k\|^2$. Note that (using Claim 11)

$$\frac{\gamma^k}{k\|u^k\|^2} = \begin{cases} O(\frac{k^{1/2}\|u^k\|^{1/2}}{k\|u^k\|^2}) = o(1) & \text{if } k\|d^k\|^3 \to \infty \\ O(1) & \text{if } k\|u^k\|^3 \to \Theta(1). \end{cases} \qquad \text{eq:fact-about-gamma} \quad (39)$$

Now, (39) and the fact that $k\|u^k\|^2 \to \infty$ from Claim 11, dividing the first-order condition by $k\|u^k\|^2 \to \infty$ yields

$$\begin{aligned} 0 \leftarrow & \sum_{j=1}^{q} \frac{\eta^k_{G,j}}{k\|u^k\|^2} \nabla_x G_j(x^k, y^k) + \sum_{\ell=1}^{p} \frac{\eta^k_{g,\ell}}{k\|u^k\|^2} \nabla_x g_\ell(x^k, y^k) - \sum_{\ell \in A_g(x^*, z^*)} \frac{\xi^k_\ell}{k\|u^k\|^2} \nabla_x g_\ell(x^k, z^*) \\ & + \sum_{i=1}^{m} \frac{\vartheta^k_i}{k||u^k||^2} \nabla_x \frac{\partial f(x^k, y^k)}{\partial y_i} + \frac{\gamma^k}{k\|u^k\|^2} [\nabla_x f(x^k, y^k) - \sum_{\ell \in A_g} \eta^k_{g,\ell} \nabla_x g_\ell(x^k, y^k) - \nabla_x f(x^k, z^*)] \end{aligned}$$

as $k \to \infty$. By Claim 9, the last term $\frac{\gamma^k}{k\|u^k\|^2}[\nabla(f(x^k, y^k) - f(x^k, z^*)) - k\sum_{\ell\in A_g} g_\ell^-(x^k, y^k)^3 \nabla g_\ell(x^k, y^k)] \to 0$. Therefore, we must have that all coefficients are zero, otherwise, again, multiplying through by the $\hat{d}$ in (Q1)–(Q8) yields a contradiction. Therefore,

$$0 \leftarrow \frac{\eta^k_{G,j}}{k\|u^k\|^2} = \frac{\min\{0, G_j(x^k, y^k)\}}{\|u^k\|^2} = \frac{(u^k)^\top \nabla^2 G_j(x^*, y^*)(u^k)}{\|u^k\|^2} \to \hat{u}^\top \nabla^2 G_j(x^*, y^*)^\top \hat{d}$$

as $k \to \infty$, where the second the second equality uses the second-order Taylor series expansion of G_j around (x^*, y^*) and utilizes the fact that $\nabla G_j(x^*, y^*)^\top \hat{d} = 0$. By uniqueness of limits, this implies that $\hat{d}^\top \nabla^2 G_j(x^*, y^*)^\top \hat{d} = 0$. By similar reasoning, we also have $\hat{d}_x^\top \nabla^2_{xx} g_\ell(x^*, z^*)^\top \hat{d}_x = 0$ for every $\ell \in A_g(x^*, z^*)$. This establishes part of the contradiction to (37b).

Finally, we turn to show the condition $\nabla g_\ell(x^*, y^*)^\top \hat{d} = 0$ for every $\ell \in A_g(x^*, y^*)$, contradicting (37d). Dividing the first-order condition $\nabla F^k(x^k, y^k, z^k) = 0$ by $k\|u^k\|^3 = \Omega(1)$ yields

$$\begin{aligned} 0 \leftarrow & \frac{\nabla_x F(x^k, y^k)}{k\|u^k\|^3} + \sum_{j=1}^{q} \frac{\eta^k_{G,j}}{k\|u^k\|^3} \nabla_x G_j(x^k, y^k) + \sum_{\ell=1}^{p} \frac{\eta^k_{g,\ell}}{k\|u^k\|^3} \nabla_x g_\ell(x^k, y^k) - \sum_{\ell \in A_g(x^*, z^*)} \frac{\xi^k_\ell}{k\|u^k\|^3} \nabla_x g_\ell(x^k, z^*) \\ & + \sum_{i=1}^{m} \frac{\vartheta^k_i}{k\|u^k\|^3} \nabla_x \frac{\partial f(x^k, y^k)}{\partial y_i} + \frac{\gamma^k}{k\|u^k\|^3} [\nabla_x f(x^k, y^k) - k \sum_{\ell=1}^{p} g_\ell^-(x^k, y^k)^3 \nabla_x g_\ell(x^k, y^k) - \nabla_x f(x^k, z^*)] \end{aligned}$$

as $k \to \infty$. Note that each coefficient of $\nabla G_j(x^k, y^k)$, $\nabla_x g_\ell(x^k, z^*)$ and $\nabla g_\ell(x^k, y^k)$ should be bounded, otherwise we get the same $0 > 0$ contradiction by multiplying through by the d that satisfies (Q1)-(Q8). In particular, since $\gamma^k \to \infty$, the coefficient $-\frac{\gamma^k}{k\|u^k\|^3} k \sum_{\ell=1}^{p} g_\ell^-(x^k, y^k)^3 = O(1)$ implies $\frac{g_\ell(x^k, y^k)^3}{\|u^k\|^3} = \left(\frac{\nabla g_\ell(x^*, y^*)^\top u^k}{\|u^k\|}\right)^3 + o(1) \to 0$, and so $\nabla g_\ell(x^*, y^*)^\top \hat{d} = 0$. Therefore, all coefficients are zero only happens when there exist some nonzero vector $\hat{d}$ satisfying (37), which is ruled out by the assumption of the theorem. □

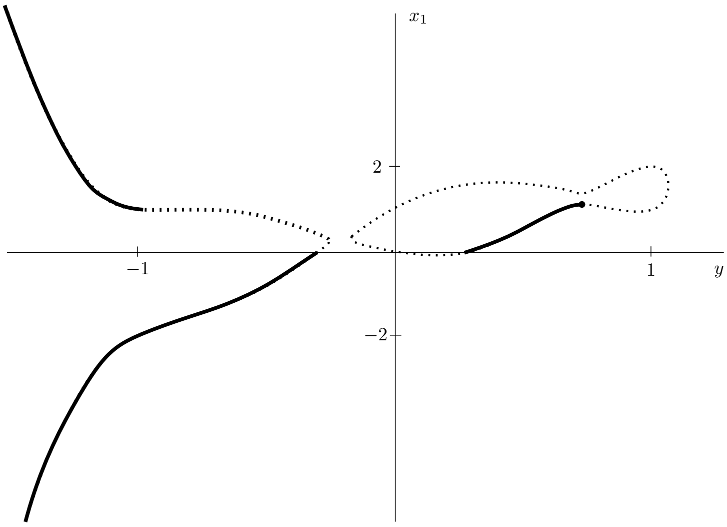

Fig. 1 The best response curve (thick line) for Example 1 in terms of x_1 and y. The dashed curve is the locus of stationary points for the lower-level problem. The horizontal axis corresponds to the follower's decision variable y and the vertical axis corresponds for the leader's first decision variable x_1. fig:best-response-curve

3 Examples

s:examples

In this section, we provide four examples of nonconvex bilevel programming problems that illustrate our constraint qualification and necessary optimality condition. The examples are chosen because either they fail the constraint qualifications of other optimality conditions found in the literature and satisfy ours, or illustrate the sparsity of our conditions compared to others in the literature. Examples 1 and 3 illustrate how our constraint qualification and optimality conditions apply where the "calmness" conditions of Ye and Zhu [16] do not. The fourth example is adapted from Ye and Zhu [16] and shows our optimality condition applies when the traditional KKT and value function approaches are invalid. Example 2 illustrates how, even when the optimality conditions of [16] apply, our optimality condition can involve fewer Lagrange multipliers.

3.1 Comparison with Theorem 2.1 of Ye and Zhu [16]

ss:compare-to-ye-unconstrained

Example 1 Solve the (BLPP) ex:theorem-2-1-fails

$$\max_{x,y} \; -(x_1 - \tfrac{6}{5})^2 + \tfrac{1}{2}(x_2 - \tfrac{y}{2})^2 - (y - \tfrac{1}{4})^2 - \tfrac{1}{2}x_2$$

$$\text{subject to } y \in \arg\max_{y \in Y} -(x_1 - 1)(x_1 - y)y - (y^2 - 1)^2(y^2 - \tfrac{1}{2})^2.$$

where $Y = [-1.5, 1.5]$

Observe that lower-level problem is in terms of x_1 and y alone. Figure 1 illustrates the best response curve; that is, the graph of $S(x_1)$. One can verify that the bilevel optimal solution is $(x^*, y^*) = (1, \frac{\sqrt{2}}{4} + \frac{1}{2}, \frac{1}{\sqrt{2}})$. First note that x_2 can be chosen in terms of x_1 and y first in an unconstrained way (x_2 will play a more substantial role in the next example). This gives the first-order condition $x_2 - \frac{y}{2} = \frac{1}{2}$. Therefore, the problem becomes

$$\min_{x_1,y}(x_1 - \tfrac{6}{5})^2 + (y - \tfrac{1}{4})^2 - \tfrac{1}{8} + \tfrac{1}{2}(\tfrac{y}{2} + \tfrac{1}{2})$$

$$s.t.\ y \in \arg\min (x_1 - 1)(x_1 - y)y + (y^2 - 1)^2(y^2 - \tfrac{1}{2})^2.$$

The first-order condition curve (which has two branches, upper and lower) are:

$$x_1^u(y) = y + \tfrac{1}{2} + \tfrac{1}{2}\sqrt{-32y^7 + 72y^5 - 52y^3 + 4y^2 + 8y + 1}$$

$$x_1^\ell(y) = y + \tfrac{1}{2} - \tfrac{1}{2}\sqrt{-32y^7 + 72y^5 - 52y^3 + 4y^2 + 8y + 1}.$$

But only part of the first-order condition curve belong to the best response. Clearly, the closest branch of the best response curve to the center $(\frac{6}{5}, \frac{3}{8})$ is on the lower branch of the first-order condition curve $\{x_1^\ell(y) : 0 \le y \le \frac{\sqrt{2}}{2}\}$. We can verify that the objective function is minimized along the lower branch at the point $y^* = \frac{\sqrt{2}}{2}$ and thus accordingly $x_1^* = x_1^\ell(y^*) = 1$. Thus, $x_2^* = \frac{\sqrt{2}}{4} + \frac{1}{2}$.

We will use alternate best response $z^* = -\frac{1}{\sqrt{2}}$. We show that under these values our constraint qualification in Theorem 4 holds. Observe that (Q1)–(Q4) are vacuous since there neither upper- nor lower-level constraints. It remains to check (Q5)–(Q9).

Observe that (Q5) amounts to $(\sqrt{2} - 1)d_{x_1} = d_y$. Condition (Q6) amounts to $d_{x_1} \le 0$. Observe that (Q7) does not constrain d_y since $\nabla_y f(x^*, y^*) = 0$, as the lower-level problem is unconstrained. We thus have tremendous freedom to choose the vector d. In fact, taking $d_{x_1} = -1$, $d_{x_2} = 1$ and $d_y = 1 - \sqrt{2}$ suffices. Finally, (Q8) is automatically satisfied since there is a single column and it is nonzero ($\nabla^2_{yy} f(x^*, y^*) = -1$). Also observe for (Q9) that $||\nabla_x f(x^*, y^*) - \nabla_x f(x^*, z^*)|| = ||(-\sqrt{2}, 0)^\top|| \ne 0$. Hence, the optimality condition in Theorem 4 applies to this problem and (33) holds. The first-order conditions have

$$\tfrac{2}{5} - \beta\sqrt{2} + \theta(\sqrt{2} - 1) = 0$$

$$-2(\tfrac{\sqrt{2}}{2} - \tfrac{1}{4}) + \tfrac{1}{4} + \beta \cdot 0 - \theta = 0$$

$$(\tfrac{\sqrt{2}}{4} + \tfrac{1}{2} - \tfrac{\sqrt{2}}{4}) - \tfrac{1}{2} + \beta \cdot 0 + \theta \cdot 0 = 0,$$

which gives unique Lagrange multipliers $\theta = \frac{3}{4} - \sqrt{2}$ and $\beta = \frac{1}{\sqrt{2}}(\frac{2}{5} - (\sqrt{2} - 1)(\sqrt{2} - \frac{3}{4})) > 0$.

We now show that this example does not satisfy the constraint qualification set in Theorem 2.1 of Ye and Zhu [16]. To be consistent with Ye and Zhu's notation, we modify the problem into minimization form with $F(x, y) = (x_1 - \frac{6}{5})^2 + (y - \frac{1}{4})^2 - \frac{1}{2}(x_2 - \frac{y}{2})^2 + \frac{1}{2}x_2$ and $f(x, y) = (x_1 - 1)(x_1 - y)y + (y^2 - 1)^2(y^2 - \frac{1}{2})^2$. The constraint qualification in Theorem 2.1 of Ye and Zhu [16] is that there exists a nonnegative real multiplier μ such that $\nabla F_\mu(x^*, y^*)^\top d \ge 0$ holds for all $d = (d_x, d_y)$ satisfying $\nabla^2_{yx} f(x^*, y^*)d_x + \nabla^2_{yy} f(x^*, y^*)d_y = 0$ where $F_\mu(x, y) = F(x, y) + \mu(f(x, y) - V(x))$ and $V(x) = \min\{f(x, y) : y \in Y\}$.

The optimal solution is $x^* = (1, \frac{\sqrt{2}}{4} + \frac{1}{2})$ and $y^* = \frac{\sqrt{2}}{2}$, $\nabla_{x_1} f(x^*, y^*) = \frac{\sqrt{2}-1}{2}$, $\nabla_{x_1} f(x^*, z^1) = 0$, $\nabla_{x_1} f(x^*, z^2) = -\frac{\sqrt{2}+1}{2}$, and $\nabla_{x_1} f(x^*, z^3) = -2$, where $(z^1, z^2, z^3) = (1, -\frac{\sqrt{2}}{2}, -1)$. Using Danskin's theorem (Proposition 2.1 in Ye and Zhu [16]), we have

$$V'(x^*, d_x) = \min\{\nabla_x f(x^*, y^*)^\top d_x, \nabla_x f(x^*, z^i)^\top d_x\} = \begin{cases} -2d_{x_1} - 0 \cdot d_{x_2} & \text{if } d_{x_1} > 0 \\ \frac{\sqrt{2}-1}{2} d_{x_1} - 0 \cdot dx_2 & \text{if } d_{x_1} \le 0, \end{cases}$$

where $\nabla_{x_2} f(x^*, y^*) = 0$ is constant. Note that the tangent plane of the first-order condition is

$$\{d \in \mathbb{R}^3 : \nabla^2_{x_1 y} f(x^*, y^*)d_{x_1} + \nabla^2_{x_2 y} f(x^*, y^*)d_{x_2} + \nabla^2_{yy} f(x^*, y^*)d_y = 0\}$$

and amounts to $-(\sqrt{2}-1)d_{x_1} + d_y = 0$, or equivalently $d_y = (\sqrt{2}-1)d_{x_1}$ and d_{x_2} is totally free.

Therefore, we have

$$\begin{aligned} F'_\mu((x^*,y^*),d) &= \nabla F(x^*,y^*)^\top d + \mu[\nabla_x f(x^*,y^*) - V'(x^*,d_x)]d_x \\ &= \begin{cases} \nabla F(x^*,y^*)^\top d + \mu\left((\frac{\sqrt{2}-1}{2} + 2)d_{x_1} + 0 \cdot d_{x_2}\right) & \text{if } d_{x_1} > 0 \\ \nabla F(x^*,y^*)^\top d + \mu\left((\frac{\sqrt{2}-1}{2} - \frac{\sqrt{2}-1}{2})d_{x_1} + 0 \cdot d_{x_2}\right) & \text{if } d_{x_1} \leq 0. \end{cases} \end{aligned}$$

Ye and Zhu's condition requires that $F'_\mu((x^*,y^*),d) \geq 0$ for any $d_y = (\sqrt{2}-1)d_{x_1}$ and d_{x_2}. Consider the direction $d_{x_1} < 0$, however we have

eq:as-we-have-shown

$$\begin{aligned} \nabla F(x^*,y^*)^\top d &= \nabla_{x_1} F(x^*,y^*)d_{x_1} + \nabla_y F(x^*,y^*)d_y + \nabla_{x_2} F(x^*,y^*)d_{x_2} \\ &= [\nabla_x F(x^*,y^*) + \nabla_y F(x^*,y^*)(\sqrt{2}-1)]d_{x_1} + \nabla_{x_2} F(x^*,y^*)d_{x_2} \qquad (40) \\ &= [-\tfrac{2}{5} + (\sqrt{2} - \tfrac{1}{2})(\sqrt{2}-1)]d_{x_1} - \tfrac{1}{2}d_{x_2} < 0 \end{aligned}$$

for large $d_{x_2} > 0$. Hence, the constraint qualification of Theorem 2.1 in [16] fails.

3.2 Sparser optimality condition in comparison to [16]

We now consider an example (which is a simplification of the previous example) where the constraint qualifications of Theorem 2.1 of [16] do hold, and compare the resulting optimality conditions. We show that our Theorem 4 can yield more parsimonious optimality conditions, involving fewer Lagrange multipliers.

ex:theorem-2-1-holds

Example 2 Solve the (BLPP)

$$\begin{aligned} &\max_{x,y} \; -(x-2)^2 - (y - \tfrac{1}{3})^2 \\ &\text{subject to } y \in \arg\max_{y \in Y} -(x-1)(x-y)y - (y^2-1)^2(y^2 - \tfrac{1}{2})^2, \end{aligned}$$

where $Y = [-1.5, 1.5]$.

The optimal solution is $x^* = 1$ and $y^* = \frac{\sqrt{2}}{2}$. Using analogous logic to the previous example, there are three alternate best responses given $x^* = 1$; $(z^1, z^2, z^3) = (1, -\frac{\sqrt{2}}{2}, -1)$. We choose $z^* = z_2 = -\frac{1}{\sqrt{2}}$ for our optimality condition.

Turning to the optimality condition of [16], we verify the constraint qualification of their Theorem 2.1 in [16]:

$$\begin{aligned} F'_\mu((x^*,y^*),d) &= \nabla F(x^*,y^*)^\top d + \mu[\nabla_x f(x^*,y^*) - V'(x^*,d_x)]d_x \\ &= \begin{cases} \nabla F(x^*,y^*)^\top d + \mu\left((\frac{\sqrt{2}-1}{2} + 2)d_x\right) & \text{if } d_x > 0 \\ \nabla F(x^*,y^*)^\top d + \mu\left((\frac{\sqrt{2}-1}{2} - \frac{\sqrt{2}-1}{2})d_x\right) & \text{if } d_x \leq 0. \end{cases} \end{aligned}$$

For the tangent plane $-(\sqrt{2}-1)d_x + d_y = 0$, for any direction $d_x < 0$, we have

$$\begin{aligned} \nabla F(x^*,y^*)^\top d &= 2(1-2)d_x + 2(\tfrac{\sqrt{2}}{2} - \frac{1}{3})d_y \\ &= d_x(-2 + 2(\tfrac{\sqrt{2}}{2} - \frac{1}{3})(\sqrt{2}-1)) > 0. \end{aligned}$$

Therefore, the constraint qualification is satisfied since there exists some $\mu \geq 0$ such that

$$F'_\mu((x^*, y^*), d) \geq 0$$

is satisfied for all direction d on the tangent plane. The optimality condition of Theorem 2.1 in [16] involves a Lagrange multiplier λ^i for each of the three best responses z^1, z^2, z^3 (for brevity we don't work out the precise details). However, using the same logic as in the previous example to verify (Q1)–(Q9), the optimality condition of our Theorem 4 applies with a single alternate best response z^*.

The optimality conditions are

$$\begin{aligned} 2 - \beta\sqrt{2} + \theta(\sqrt{2} - 1) = 0 \\ -2(\tfrac{\sqrt{2}}{2} - \tfrac{1}{3}) + \beta \cdot 0 - \theta = 0, \end{aligned}$$

which yields $\theta = 2(\frac{\sqrt{2}}{2} - \frac{1}{3})$ and $\beta = \frac{2-(\sqrt{2}-1)(\sqrt{2}-\frac{2}{3})}{\sqrt{2}} > 0$.

3.3 Comparison with Theorem 3.2 of Ye and Zhu [16]

ss:compare-to-ye-constrained

Our third example provides a case where our constraint qualifications hold, but the calmness conditions of Theorem 3.2 in [16] fail to hold. This example is a constrained version of Example 1 with a single upper-level constraint.

ex:theorem-3-2-fails

Example 3 Consider the following bilevel program:

$$\begin{aligned} \max_{x,y} \quad & -(x_1 - \tfrac{6}{5})^2 + \tfrac{1}{2}(x_2 - \tfrac{y}{2})^2 - (y - \tfrac{1}{4})^2 - \tfrac{1}{2}x_2 \\ \text{subject to} \quad & \tfrac{\sqrt{2}-1}{2}x_1 - y + 1/2 \geq 0 \\ & y \in \arg\max_{y \in Y} -(x_1 - 1)(x_1 - y)y - (y^2 - 1)^2(y^2 - \tfrac{1}{2})^2. \end{aligned}$$

Since the solution $(x^*, y^*) = (1, \frac{\sqrt{2}}{4} + \frac{1}{2}, \frac{1}{\sqrt{2}})$ satisfies the upper-level constraint $G(x, y) = \frac{\sqrt{2}-1}{2}x_1 - y + 1/2 = 0$ it remains a bilevel optimal solution. Constraint qualification (Q2) requires that $\frac{\sqrt{2}-1}{2}d_{x_1} > d_y$. Carrying over from Example 1 we also require that $(\sqrt{2} - 1)d_{x_1} = d_y$ and $d_{x_1} \leq 0$. Taking $d_{x_1} = -1$, $d_{x_2} = 0$ and $d_y = 1 - \sqrt{2}$ suffices.

We now show that the constraint qualifications in Theorem 3.2 of [16] fail. Since there is an upper-level constraint, Theorem 2.1 of [16] no longer applies as in the previous examples.

We again modify the problem from maximization form to minimization form (as in Example 1) and add the corresponding constraint: $G(x, y) = -\frac{\sqrt{2}-1}{2}x_1 + y - \frac{1}{2} \leq 0$. Based on our calculation in Example 1, $\nabla_{x_1} f(x^*, y^*) = \frac{\sqrt{2}-1}{2}$, $\nabla_{x_1} f(x^*, z^2) = -\frac{1+\sqrt{2}}{2}$, and $\nabla_{x_1} f(x^*, z^3) = -2$, where $z^1 = 1$ is now ruled out by the upper-level constraint. This means, $W(x^*) = \{\nabla_x f(x^*, y') : y' \in S(x^*)\} = \{(-2, 0)^\top, (-\frac{1+\sqrt{2}}{2}, 0)^\top, (\frac{\sqrt{2}-1}{2}, 0)^\top\}$, where $W(x^*)$ is as defined in equation (1.4) of [16]. Therefore, the MPEC linearization cone (Definition 3.4 in [16]) is

$$\begin{aligned} \mathcal{L}^{\text{MPEC}}(x^*, y^*) &= \{d \in \mathbb{R}^3 : \nabla(\nabla_y f(x^*, y^*))d = 0, \nabla G(x^*, y^*)d \leq 0\} \\ &= \{d \in \mathbb{R}^3 : -(\sqrt{2} - 1)d_{x_1} + d_y = 0, -\tfrac{\sqrt{2}-1}{2}d_{x_1} + d_y \leq 0\}. \end{aligned}$$

Note that $d \in \mathcal{L}^{MPEC}(x^*, y^*)$ implies that only $d_{x_1} \leq 0$ is feasible, by $(\sqrt{2}-1)d_x = d_y \leq \frac{\sqrt{2}-1}{2} d_{x_1}$. MPEC-weak calmness at (x^*, y^*) with modulus $\mu > 0$ requires

$$[\nabla F(x^*, y^*) + \mu \nabla f(x^*, y^*)]^\top d - \mu \min_{\xi \in W(x^*)} \xi^\top d_x \geq 0 \text{ for all } d \in \mathcal{L}^{\text{MPEC}}(x^*, y^*).$$

By $d_y = (\sqrt{2}-1)d_{x_1}$, the above inequality is equivalent to

$$\nabla F(x^*, y^*)^\top d + \mu \tfrac{\sqrt{2}-1}{2} d_{x_1} - \mu \tfrac{\sqrt{2}-1}{2} d_{x_1} \geq 0 \text{ for any } d \in \mathcal{L}^{\text{MPEC}}(x^*, y^*).$$

which amounts to $\nabla_{x_1} F(x^*, y^*) d_{x_1} + \nabla_y F(x^*, y^*) d_y + \nabla_{x_2} F(x^*, y^*) d_{x_2} \geq 0$ since $d_{x_1} \leq 0$ when $d \in \mathcal{L}^{\text{MPEC}}(x^*, y^*)$ and $\xi = (\frac{\sqrt{2}-1}{2}, 0)^\top$ is chosen from $W(x^*)$. However, as we have shown in (40) for a given $d_{x_1} < 0$,

$$\begin{aligned}
\nabla F(x^*, y^*)^\top d &= \nabla_{x_1} F(x^*, y^*) d_{x_1} + \nabla_y F(x^*, y^*) d_y + \nabla_{x_2} F(x^*, y^*) d_{x_2} \\
&= [\nabla_x F(x^*, y^*) + \nabla_y F(x^*, y^*)(\sqrt{2}-1)] d_{x_1} + \nabla_{x_2} F(x^*, y^*) d_{x_2} \\
&= [-\tfrac{2}{5} + (\sqrt{2} - \tfrac{1}{2})(\sqrt{2}-1)] d_{x_1} - \tfrac{1}{2} d_{x_2} < 0
\end{aligned}$$

for d_{x_2} sufficiently large. Therefore, MPEC weak calmness is not satisfied. Since there is no lower-level constraint, MPEC-weakly calm is the same as weakly calm and so the constraint qualification in Theorem 3.2 in Ye and Zhu [16] do not apply to this instance.

3.4 Comparison with classical conditions

ss:compare-to-classical

Our final example is adapted from [16] (Example 4.3) and converted into our format (that is, a maximization problem). Ye and Zhu used this example to illustrate how the classic KKT approach and value functions approaches may fail but the weak calmness condition of [16] nonetheless hold. We show that this problem also satisfies our constraint qualification and yields the same optimality condition.

ex:others-fail

Example 4 Consider the following bilevel program:

$$\max_{x,y} -(x-0.5)^2 - (y-2)^2$$

$$\text{subject to } y \in S(x) := \arg \max_{y \in [-4,4]} \{3y - y^3 : y \geq x - 3\}.$$

In terms of our notation, we have $F(x, y) = -(x-0.5)^2 - (y-2)^2$, $f(x, y) = 3y - y^3$, $g(x, y) = -x + y + 3$ and $Y = [-4, 4]$. Note that G is not present. In [16], Ye and Zhu argue that $x^* = y^* = 1$ is bilevel optimal solution. There is a unique alternate best response $z^* = -2$. Observe that both y^* and z^* are in the interior of Y.

We now examine the constraint qualifications of Theorems 4 and 5. First, observe that (Q9) does not hold and so Theorem 5 applies. It remains to check (Q1) to (Q8). To see that (Q1) holds, any $d_z < 0$ suffices. Condition (Q2) holds vacuously. Since $g_\ell(x^*, y^*) \neq 0$, condition (Q3) also holds vacuously.

The remaining conditions concern simultaneously satisfies some inequalities involving d_x and d_y. Condition (Q4) restricts $d_x > 0$ since $g_x(x^*, z^*) = -1$. Condition (Q5) amounts to $-6d_y = 0$ (since $\nabla^2_{xy} f(x, y) = 0$ and $\nabla^2_{yy} f(x^*, y^*) = -6$) and so $d_y = 0$. Condition (Q6) holds trivially since $\nabla f(x^*, y^*) = (0, 0)^\top$. Condition (Q7) holds vacuously since $f_y(x^*, y^*) = 0$. Together this implies that d satisfying $d_x > 0$ and $d_y = 0$ works for (Q2) to (Q8).

Finally, we show there does not exist a $\hat{d}$ such that (37) holds. Indeed, from (37b) we have $\nabla_x g(x^*, z^*)^\top \hat{d}_x = 0$, which implies that $-\hat{d}_x = 0$ or $\hat{d}_x = 0$. Moreover, from (37c) we know $\left(\nabla \frac{\partial f(x^*,y^*)}{\partial y_i}\right)^\top \hat{d} = 0$, which implies $\nabla^2_{yy} f(x^*, y^*)\hat{d}_y = -6\hat{d}_y = 0$ or $\hat{d}_y = 0$. This implies $\hat{d}_x = \hat{d}_y = 0$ and so there does not exist a nonzero $\hat{d}$ such that (37) holds.

By Theorem 5, this implies that optimality conditions (33) and complementary slackness conditions (25) hold. Moreover, from (33a) we have: $-1 = -\kappa + \lambda_g + \rho_\ell$ and and from (33b) $-2 = \lambda_g + \rho_\ell - 6\theta$. Any multipliers satisfying these inequalities provide a characterization of optimality.

References

1. D.P. Bertsekas, *Nonlinear programming*, Athena Scientific, 1999.
2. B. Colson, P. Marcotte, and G. Savard, *An overview of bilevel optimization*, Annals of Operations Research **153** (2007), no. 1, 235–56.
3. S. Dempe, *Foundations of bilevel programming*, Springer, 2002.
4. S. Dempe, J. Dutta, and B.S. Mordukhovich, *New necessary optimality conditions in optimistic bilevel programming*, Optimization **56** (2007), no. 5-6, 577–604.
5. S. Dempe and A.B. Zemkoho, *The generalized Mangasarian-Fromowitz constraint qualification and optimality conditions for bilevel programs*, Journal of Optimization Theory and Applications **148** (2011), no. 1, 46–68.
6. ———, *The bilevel programming problem: reformulations, constraint qualifications and optimality conditions*, Mathematical Programming (2013), 1–27.
7. Y. Ishizuka and E. Aiyoshi, *Double penalty method for bilevel optimization problems*, Annals of Operations Research **34** (1992), no. 1, 73–88.
8. R. Ke and C.T. Ryan, *A general solution method for moral hazard problems*, to appear in Theoretical Economics (2018).
9. ———, *Monotonicity of optimal contracts without the first-order approach*, to appear in Operations Research (2018).
10. G.S. Liu, J.Y. Han, and J.Z. Zhang, *Exact penalty functions for convex bilevel programming problems*, Journal of Optimization Theory and Applications **110** (2001), no. 3, 621–643.
11. P. Marcotte and D.L. Zhu, *Exact and inexact penalty methods for the generalized bilevel programming problem*, Mathematical Programming **74** (1996), no. 2, 141–157.
12. Z. Meng, C. Dang, R. Shen, and M. Jiang, *An objective penalty function of bilevel programming*, Journal of Optimization Theory and Applications **153** (2012), no. 2, 377–387.
13. P. Milgrom and I. Segal, *Envelope theorems for arbitrary choice sets*, Econometrica **70** (2002), no. 2, 583–601.
14. J.A. Mirrlees, *The theory of moral hazard and unobservable behaviour: Part I*, The Review of Economic Studies **66** (1999), no. 1, 3–21.
15. J.J. Ye and D.L. Zhu, *Optimality conditions for bilevel programming problems*, Optimization **33** (1995), no. 1, 9–27.
16. ———, *New necessary optimality conditions for bilevel programs by combining the MPEC and value function approaches*, SIAM Journal on Optimization **20** (2010), no. 4, 1885.

A Appendix: Some technical proofs

A.1 Proof of Claim 2

ss:app-speed-of-convergence

We first establish the following simpler lemma.

lemma:convergence-rate-lower-level

Lemma 5 *If (Q1) holds then* $\hat{f}(k) - f^* = O(k^{-1/3})$ *where, for constant* $\alpha > 0$, $f^k(z) := f(x^*, z) - \frac{k}{4}\sum_{\ell=1}^{p}(g_\ell^-(x^*, z))^4 - \frac{\alpha}{2}\|z - z^*\|^2$ *is an essentially standard penalty function for (LLP),*[7] $z^k \in \arg\max_{z \in Y} f^k(z)$, *and* $\hat{f}(k) := \max_{z \in Y} f^k(z) = f^k(z^k)$. *In other words,* f^k *is an exact penalty function for all choice of* α.

[7] In [1], the penalty function $f^k(z) := f(x^*, z) - \frac{k}{2}\sum_{\ell=1}^{p}(g_\ell^-(x^*, z))^2 - \frac{\alpha}{2}\|z - z^*\|^2$ would be used; that is, we change the power from 2 to 4.

Proof We need to estimate the rate at which $\lim_{k\to\infty} \hat{f}(k) - f^*$ goes to 0. As a first step we examine $\lim_{k\to\infty} k^{1/3}(\hat{f}(k) - f^*)$. Observe that

eq:claim-1

$$\lim_{k\to\infty} k^{1/3}(\hat{f}(k) - f^*) = \lim_{k\to\infty} \frac{\hat{f}(k) - f^*}{1/k^{1/3}}. \tag{41}$$

We work to establish the numerator is differentiable in k in order to apply L'Hôpital's rule with respect to k to the latter expression. To prove differentiability of the numerator, we first observe that $\hat{f}(k)$ is decreasing in k. Moreover, by exactness of the penalty function of the lower level problem (see Section 3.3 in [1]), $\hat{f}(k) \to f^*$ as $k \to \infty$ and thus $-\infty < \hat{f}(k) < \infty$ for all k since $\hat{f}(k)$ must approach f^* from above. Moreover, $\hat{f}$ is a *convex* function of k and so is differentiable on its domain $\mathbb{R}$. Thus, $\hat{f}$ is a differentiable function of k.

Now applying L'Hôpital's rule to the right-hand side of (41) yields:

eq:claim-2

$$\lim_{k\to\infty} k^{1/3}(\hat{f}(k) - f^*) = \lim_{k\to\infty} \frac{-\frac{1}{4}\sum_{\ell=1}^{p} g_\ell^-(x^*, z^k)^4}{-\frac{1}{3}k^{-4/3}} \tag{42}$$

since

$$\begin{aligned}
\tfrac{d}{dk}(\hat{f}(k) - f^*) &= \tfrac{d}{dk}\left[\max_{z\in Y} f(x^*, z) - \tfrac{k}{4}\sum_{\ell=1}^{p}(g_\ell^-(x^*, z))^4 - \tfrac{\alpha_0}{2}\|z - z^*\|^2\right] \\
&= \tfrac{\partial}{\partial k}\left[f(x^*, z^k) - \tfrac{k}{4}\sum_{\ell=1}^{p}(g_\ell^-(x^*, z^k))^4 - \tfrac{\alpha_0}{2}\|z^k - z^*\|^2\right] \\
&= -\tfrac{1}{4}\sum_{\ell=1}^{p}(g_\ell^-(x^*, z^k))^4,
\end{aligned}$$

where the first equality observes that f^* is a constant with respect to k, the second equality writes out $\hat{f}(k)$, the third equality leverages the Envelope Theorem, and the final equality takes the coefficient off the linear term in k in the previous equality.

From (42) the next step is to examine the asymptotic growth rate of $g_\ell^-(x^*, z^k)$ in terms of k. We now leverage the assumption that z^* satisfies the MFCQ and claim it implies that $kg_\ell^-(x^*, z^k)^3$ is bounded. Suppose otherwise, then the Fritz-John coefficient $\mu_0 = 0$. However, under (Q1) we know $\mu_0 \neq 0$, a contradiction. Since $kg_\ell^-(x^*, z^k)^3$ is bounded, this means $g_\ell^-(x^*, z^k)$ is $O(k^{-1/3})$. This, in turn, implies in (42) that $\sum_{\ell=1}^{p} g_\ell^-(x^*, z^k)^4$ is $O(k^{-4/3})$. That is, from (42) we have $k^{1/3}(\hat{f}(k) - f^*)$ is $O(k^{4/3}k^{-4/3}) = O(1)$ and so $\hat{f}(k) - f^*$ is $O(k^{-1/3})$, as required. □

cor:delta-differentiable

Corollary 2 *The function $\Delta(k)$ defined in (7) is convex, decreasing and converges in value to 0 as $k \to \infty$.*

The proof of this corollary follows identical logic to the proof of Lemma 5. Lemma 5 can then be strengthened to establish Claim 2. First, we choose $\alpha_3 > 0$ is sufficiently large such that MFCQ holds within neighborhood $z \in \bar{B}_{\sqrt{2/\alpha_3}}(z^*) \subset Y$. Next, define $\rho(k, \alpha) = \max_{z\in Y}\left\{f(x^*, z) - \frac{k}{4}\sum_{\ell=1}^{p}\left(g_\ell^-(x^*, z)\right)^4 - \frac{\alpha}{2}\|z - z^*\|^2\right\}$, which is the maximum of z over a penalty function of the lower level problem with constant α (and so by Lemma 5 is an exact penalty function) and

$$S^k(x^*; \alpha) = \arg\max_{z\in Y}\left\{f(x^*, z) - \tfrac{k}{4}\sum_{\ell=1}^{p}\left(g_\ell^-(x^*, z)\right)^4 - \tfrac{\alpha}{2}\|z - z^*\|^2\right\}.$$

Below we show the choice of α may depend on k. Then for any $z_{k,\alpha} \in S^k(x^*; \alpha)$, we claim that

distance

$$\|z_{k,\alpha} - z^*\|^2 \leq \tfrac{2\Delta(k)}{\alpha}, \tag{43}$$

where $\Delta(k)$ is defined in (7). Indeed, by definition, note that $\rho(k, \alpha) \geq f(x^*, z^*)$. Thus

$$f(x^*, z_{k,\alpha}) - \tfrac{k}{4}\sum_{\ell=1}^{p}\left(g_\ell^-(x^*, z_{k,\alpha})\right)^4 - \tfrac{\alpha}{2}\|z_{k,\alpha} - z^*\|^2 \geq f(x^*, z^*),$$

which implies that for $k \geq k_0$ sufficiently large

$$\begin{aligned}
\tfrac{\alpha}{2}\|z_{k,\alpha} - z^*\|^2 &\leq f(x^*, z_{k,\alpha}) - \tfrac{k}{4}\sum_{\ell=1}^{p}\left(g_\ell^-(x^*, z_{k,\alpha})\right)^4 - f(x^*, z^*) \\
&\leq f(x^*, z_{k,\alpha}) - \tfrac{k}{4}2^{-9}\sum_{\ell=1}^{p}\left(g_\ell^-(x^*, z_{k,\alpha})\right)^{10} - f(x^*, z^*) \leq \Delta(k),
\end{aligned}$$

where the second inequality uses the fact that $g_\ell^-(x^*, z_{k,\alpha})^4 \to 0$ implies $g_\ell^-(x^*, z_{k,\alpha})^4 \geq g_\ell^-(x^*, z_{k,\alpha})^{10}$. Note that the constant 2^{-9} is not relevant to the argument here but will be used later.

On the other hand, by Assumption (Q1), there exists a vector $d \in \mathbb{R}^m$ such that $\nabla_y g_\ell(x^*, z^*) \cdot d > 0$ for all $\ell \in A_g(x^*, z^*)$. Hence by the continuity of $\nabla_y g_\ell$ and the compactness of a small neighborhood $\mathcal{N}_\delta(z^*) \subset Y$, there are positive constants c_0 and δ such that for all $\|z - z^*\| \leq \delta$,

MFCQ-neighborhood

$$\nabla_y g_\ell(x^*, z) \cdot d \geq c_0 > 0, \; \forall \, \ell \in A_g(x^*, z^*). \tag{44}$$

Taking α as a function of k as $\alpha = \alpha(k) = \frac{2\varDelta(k)}{\delta^2} \leq \frac{2\varDelta(k_0)}{\delta^2}$, for any $k \geq k_0$, then $\alpha(k)$ is decreasing and convex by the property of $\varDelta(k)$. Moreover, by using (43), we have $\|z_{k,\alpha(k)} - z^*\| \leq \delta$. Hence (44) holds for any $z = z_{k,\alpha(k)}$. Now we claim that for all $k \geq k_0$,

estimate1

$$\left|g_\ell^-(x^*, z_{k,\alpha(k)})\right| \leq Ck^{-1/3}, \tag{45}$$

where C is a positive constant independent of k. To prove (45), by the first-order condition of the maximization problem defining $\rho(k, \alpha)$, we have

$$\nabla_y f(x^*, z_{k,\alpha(k)}) - k \sum_{\ell=1}^p \left(g_\ell^-(x^*, z_{k,\alpha(k)})\right)^3 \nabla_y g_\ell^-(x^*, z_{k,\alpha(k)}) - \alpha(z_{k,\alpha(k)} - z^*) = 0.$$

Multiplying the above equality by d (and rearranging a bit), we get

$$k \sum_{\ell=1}^p \left(g_\ell^-(x^*, z_{k,\alpha(k)})\right)^3 \nabla_y g_\ell^-(x^*, z_{k,\alpha(k)}) \cdot d = \nabla_y f(x^*, z_{k,\alpha(k)}) \cdot d - \alpha(z_{k,\alpha(k)} - z^*) \cdot d,$$

which implies that $\left|k \sum_{\ell=1}^p \left(g_\ell^-(x^*, z_{k,\alpha(k)})\right)^3\right| \leq C$, otherwise the MFCQ for the lower level problem at $z_{k,\alpha(k)}$ is violated. Hence (45) holds.

Now we claim that there exists a constant C such that for $k \geq k_0$, $\rho(k, \alpha(k)) - f^* \leq \frac{3C^4}{4} k^{-1/3}$. Note that the penalty function $f(x^*, z) - \frac{k}{4} \sum_{\ell=1}^p \left(g_\ell^-(x^*, z)\right)^4 - \frac{\alpha(k)}{2} \|z - z^*\|^2$ is differentiable with respect to k and the derivative

$$-\tfrac{1}{4} \sum_{\ell=1}^p \left(g_\ell^-(x^*, z)\right)^4 - \tfrac{\alpha'(k)}{2} \|z - z^*\|^2$$

is bounded for any z since g_ℓ^- is continuous and $z \in Y$ comes from a compact set. Then by the envelope theorem (c.f., Theorem 2 in [13]), the fundamental theorem of integration applies to the value function $\rho(k, \alpha(k))$. Then,

$$\begin{aligned}
\rho(k, \alpha(k)) - f^* &= -\left(\lim_{r \to \infty} \rho(r, \alpha(r)) - \rho(k, \alpha(k))\right) \\
&= -\int_k^\infty [\tfrac{d\rho(r,\alpha(r))}{dr}] dr \\
&= -\int_k^\infty \left[-\tfrac{1}{4} \sum_{\ell=1}^p \left(g_\ell^-(x^*, z_{r,\alpha(r)})\right)^4 - \tfrac{\alpha'(r)}{2} \|z_{r,\alpha(r)} - z^*\|^2\right] dr \\
&\leq \int_k^\infty \tfrac{C^4}{4} r^{-4/3} dr \\
&\leq \tfrac{3C^4}{4} k^{-1/3} = O(k^{-1/3}),
\end{aligned}$$

where the first step uses Lemma 5 to imply $\lim_{r \to \infty} \rho(r, \alpha(r)) = f^*$, the second step is by the fundamental theorem of integration, the third and fourth step uses the envelope theorem. ◁

A.2 Proof of Claim 3

ss:app-xfinity-works-with-ystar

The proof is by contradiction. If (x_∞, y_∞) is not bilevel feasible then for sufficiently large k

eq:drown-out-y

$$\max_{\hat{z}} \left[f(x^k, \hat{z}) - \frac{k}{4} \sum_{\ell=1}^p (g_\ell^-(x^k, \hat{z}))^4 - \frac{\alpha_3}{2} \varDelta(k) \|\hat{z} - z^*\|^2\right] > f(x^k, y^k) - \frac{k}{4} \sum_{\ell=1}^p (g_\ell^-(x^k, y^k))^4. \tag{46}$$

Indeed, as k gets large $f(x^k, y^k)$ gets closer to $f(x_\infty, y_\infty)$ and $\frac{k}{4} \sum_{\ell=1}^p (g_\ell^-(x^k, \hat{z}))^4$ gets arbitrarily close to 0, otherwise exactness again delivers a contradiction. Also, for k sufficiently large it is clear that $\hat{z}$ is chosen in the arg max of the left-hand side of (46) so that $g_\ell^-(x^k, \hat{z}) \to 0$ for all ℓ, otherwise the value of the problem gets driven to $-\infty$ as $k \to \infty$. In particular, since (x_∞, y_∞) is not bilevel feasible there exists a z' such that $g(x_\infty, z') \geq 0$ and $f(x_\infty, z') > f(x_\infty, y_\infty)$. The choices of $\hat{z}$ can approach such a z' and so eventually (46) holds. In particular, let $\hat{z}^k \in \arg\max_{\hat{z}}[f(x^k, \hat{z}) - \frac{k}{4}\sum_{\ell=1}^p (g_\ell^-(x^k, \hat{z}))^4 - \frac{\alpha_3}{2}\Delta(k)\|\hat{z} - z^*\|^2]$. Under this choice

$$\begin{aligned}\lim_{k\to\infty} \max_{x,y} \min_z F^k(x, y, z) &= \lim_{k\to\infty} \min_z F^k(x^k, y^k, z) \leq \lim_{k\to\infty} F^k(x^k, y^k, \hat{z}^k)\\ &\leq \lim_{k\to\infty} \Bigg(F(x^k, y^k) + \tfrac{k^{3/4}}{2}\|\hat{z}^k - z^*\|^2 - \tfrac{k}{4}\Big(\min\Big\{0, f(x^k, y^k) - \tfrac{k}{4}\sum_{\ell=1}^p (g_\ell^-(x^k, y^k))^4\\ &\qquad - [f(x^k, \hat{z}^k) - \tfrac{k}{4}\sum_{\ell=1}^p (g_\ell^-(x^k, \hat{z}^k))^4 - \tfrac{\alpha_3}{2}\Delta(k)\|\hat{z}^k - z^*\|^2]\Big\}\Big)^4\Bigg),\end{aligned} \tag{47}$$

eq:this-part-here

where the first equality is by definition of (x^k, y^k) and the first inequality comes from the minimization over z. The second inequality writes out the definition of $F^k(x^k, y^k, \hat{z}^k)$ from (6) dropping negative penalty terms. Now, from (46) we know that for sufficiently large k the "min" in (47) is strictly less than 0. It follows that the last term in (47) diverges to $-\infty$ as $k \to \infty$ at a linear rate in k. The second term grows to infinity at a sublinear rate in k. This implies that (47) diverges to $-\infty$, in contradiction of exactness.

A.3 Proof of Claim 4

ss:app-z-infinity-in-interior

Suppose that $z_\infty \notin S(x^*)$. Since $S(x^*)$ is a closed set (by the Theorem of Maximum), for k sufficiently large, $z^k \notin S(x^*)$. Hence, for k sufficiently large $f(x^k, y^k) - \frac{k}{4}\sum_{\ell=1}^p (g_\ell^-(x^k, y^k))^4 > f(x^k, z^k) - \frac{k}{4}\sum_{\ell=1}^p (g_\ell^-(x^k, z^k))^4$ since the left-hand side will converge to $f(x^*, y^*)$ and the right-hand side will converge to $f(x^*, z_\infty)$ where z_∞ is feasible for the lower level problem given x^*. This implies that $A^k(z^k) = 0$ for k sufficiently large, and so penalty term (viii) disappears for k sufficiently large. Hence, z^k minimizes the sum of terms (vi) and (vii). To ease notation, let $\tau = ||z - z^*||^2$ (and specifically $\tau^k = ||z^k - z^*||^2$), $d^k = (x^k, y^k)$ and $d^* = (x^*, y^*)$. Then we can express the minimization of terms (vi) and (vii) (dividing through by the constant $k^{3/4}$) as minimizing the function

$$\phi^k(\tau; d) = \tau(\tfrac{1}{2} - \tfrac{k^{1/4}}{4}||d - d^*||^2\tau). \tag{48}$$

eq:function-to-minimize

Observe that $\phi^k(\tau; d)$ is concave and so the minimization of $\phi^k(\tau; d)$ occurs at the boundaries of possible choices of τ, namely 0 and

$$\begin{aligned}\bar{\tau} &= \max_{z\in Y} ||z - z^*||^2 \text{ and}\\ \bar{z} &\in \arg\max_{z\in Y} ||z - z^*||^2.\end{aligned}$$

Now, it must be that $\phi^k(\tau^k; d^k) \leq 0$, otherwise we can set $z^k = z^*$ to make $-\frac{k}{4}(A^k(z^*))^4 + k^{3/4}\phi^k(0; d^k) \leq 0 < k^{3/4}\phi^k(\tau^k; d^k)$, a contradiction of that fact that z^k is a minimizer. From (48), this implies

$$\tfrac{k^{1/4}}{4}||d^k - d^*||^2\tau^k \geq 1/2. \tag{49}$$

eq:converge-to-half

By the Exactness Lemma, and the fact shown above that term (viii) converges to 0, we know $\phi^k(\tau^k; d^k) \to 0$. Restrict to a subsequence of the k so that $\frac{k^{1/4}}{4}||d^k - d^*||^2$ is monotone (abusing notating, again denote this subsequence by k). From (49), there are two cases to consider: $\frac{k^{1/4}}{4}||d^k - d^*||^2\tau^k$ converges to 1/2 from above (but not eventually equal to 1/2) or $\frac{k^{1/4}}{4}||d^k - d^*||^2\tau^k = 1/2$ for all k sufficiently large.

Case 1: $\frac{k^{1/4}}{4}||d^k - d^*||^2\tau^k \to 1/2$ (but not equal to 1/2) for all k sufficiently large.

For the first case, $\tau^k = \bar{\tau}$ is the unique minimizer of (48) for k sufficiently large. By the Theorem of Maximum, $\phi^k(\tau^k; d) = \phi^k(\bar{\tau}; d)$ is continuous and differentiable in d around a sufficiently small neighborhood of $d^k = (x^k, y^k)$. We obtain a contradiction of the definition of the optimality of (x^k, y^k). Perturb (x^k, y^k) by ϵ (chosen sufficiently small) in direction $\frac{(d^k - d^*)}{||d^k - d^*||}$. The change in the value of the penalty function $F^k(x^k, y^k, z^k)$ is of order

$$\nabla_{x,y}F^k(x^k, y^k, \bar{z})^\top \tfrac{(d^k - d^*)}{||d^k - d^*||}\epsilon \tag{50}$$

eq:order-to-make-big

since $\phi^k(\tau^k; d) = \phi^k(\bar{\tau}; d)$ is continuous and differentiable by the envelope theorem. The rest of the proof shows that ϵ can be chosen so that the penalty function $F^k(x^k, y^k, z^k)$ can be improved by going in direction $\frac{(d^k - d^*)}{||d^k - d^*||}\epsilon$, which violates the optimality of (x^k, y^k).

In the next step we analyze the expression in (50) term-by-term using the definition of $F^k(x, y, z)$ in (6). We have already argued that (thankfully) term (viii) is 0 for k sufficiently large. Also observe that $\nabla F(x^k, y^k)^\top \frac{(d^k - d^*)}{||d^k - d^*||} = O(1)$ since the gradient of F is bounded over the compact feasible region $X \times Y$. Clearly, the contribution of terms (iv) and (v) is negligible, and terms (vi) and (vii) is captured by analysis of ϕ. The following subclaim is relevant to penalty terms (i)–(iii), (ix) and (x).

claim:terms-go-away

Subclaim 1

eq:first-one

$$
\begin{aligned}
kG_j^-(x^k, y^k)[\nabla G_j(x^k, y^k)]^\top \tfrac{(d^k - d^*)}{||d^k - d^*||} &= o(k||d^k - d^*||^7) \textit{ for all } j = 1, \ldots, q, \qquad (51)\\
kg_\ell^-(x^k, y^k)[\nabla g_\ell(x^k, y^k)]^\top \tfrac{(d^k - d^*)}{||d^k - d^*||} &= o(k||d^k - d^*||^7) \textit{ for all } \ell = 1, \ldots, p,\\
kg_\ell^+(x^k, z^*)[\nabla g_\ell(x^k, z^*)]^\top \tfrac{(d^k - d^*)}{||d^k - d^*||} &= o(k||d^k - d^*||^7) \textit{ for all } \ell = 1, \ldots, p,\\
k\left(\min\{0, f(x^k, y^k) - \tfrac{k}{4}\sum_{\ell=1}^p g_\ell^-(x^k, y^k)^4 - f(x^k, z^*)\}\right) \times &\\
\left(\nabla_x f(x^k, y^k) - k\sum_{\ell=1}^p g_\ell^-(x^k, y^k)^3 \nabla_x g_\ell(x^k, y^k) - \nabla_x f(x^k, z^*)\right)^\top \tfrac{(d^k - d^*)}{||d^k - d^*||} &= o(k||d^k - d^*||^7),\\
k(\nabla_y f(x^k, y^k) - \nabla_y f(x^*, y^*))^\top \tfrac{(d^k - d^*)}{||d^k - d^*||} &= o(k||d^k - d^*||^7)
\end{aligned}
$$

The proof of each part of the subclaim follows a similar trajectory. We focus attention on the first statement and suppress details for the remaining. By feasibility of (x^*, y^*), we know $G_j(x^*, y^*) \geq 0$. If $G_j(x^*, y^*) > 0$ then for k sufficiently large the left-hand side of (51) is 0 and certainly $o(k||d^k - d^*||^7)$. So, suppose $G_j(x^*, y^*) = 0$. By the Exactness Lemma, we know term (i) converges to 0 and thus,

eq:to-divide

$$k^{1/2} G_j^-(x^k, y^k) \to 0 \tag{52}$$

as $k \to \infty$. From the condition of this case, we know $k^{1/2}||d^k - d^*||^4 = \Theta(1)$. Therefore, dividing (52) through by $k^{1/2}||d^k - d^*||^4$ yields

eq:when-you-divide

$$G_j^-(x^k, y^k) = o(||d^k - d^*||^4). \tag{53}$$

At the same time, taking a Taylor expansion of $G_j(x^k, y^k)$ centered at (x^*, y^*) yields

eq:second-piece

$$G_j(x^k, y^k) = \nabla_j G_j(x^k, y^k)^\top (d^k - d^*) + \text{h.o.t.} \tag{54}$$

since $G_j(x^*, y^*) = 0$ and so $\nabla_j G_j(x^k, y^k)^\top (d^k - d^*) = o(||d^k - d^*||^4)$. Putting (53) and (54) together yields

$$kG_j^-(x^k, y^k) \nabla G_j(x^k, y^k)^\top \tfrac{(d^k - d^*)}{||d^k - d^*||} = o(k||d^k - d^*||^7),$$

establishing (51).

Finally, we examine the contribution of terms (vi) and (vii) to (50). Under the condition of Case 1, we know $k^{1/4}||d^k - d^*||^2 = \Theta(1)$. Also, since $||z_k - z_\infty||^4$ is a constant at $\bar{\tau}^4$ this implies that $k||d^k - d^*||||z_k - z_\infty||^4$ diverges to ∞ and dominates all terms in (51). However, $k||d^k - d^*||||z_k - z_\infty||^4$ is precisely the inner product of the derivative of terms (vi) and (vii) with respect to (x, y) with the vector $\frac{(d^k - d^*)}{||d^k - d^*||}$. Taken together, this implies that the expression in (50), $\nabla_{x,y} F^k(x^k, y^k, \bar{z})^\top \frac{(d^k - d^*)}{||d^k - d^*||}\epsilon$ diverges to ∞ as $k \to \infty$ since ϵ can be chosen to be either positive or negative. This violates the definition of (x^k, y^k) as a maximizer for k sufficiently large, leading to a contradiction. This completes Case 1.

Case 2: $\frac{k^{1/4}}{4}||d^k - d^*||^2 \tau^k = 1/2$ for all k sufficiently large.

The approach is to mimic the ideas of Case 1, but there is a complication. In Case 1 we were able to use the standard Theorem of Maximum to argue that

eq:to-apply-theorem-of-maximum

$$\min_{z \in Y} F^k(d, z) := \min_{z \in Y} k^{3/4} \phi^k(||z - z^*||; d) - kA^k(z)^4 \tag{55}$$

is a differentiable function of d sufficiently close to d^k for all k and then using an Envelope-like theorem to yield the desired conclusion. However, this relied on the fact that there is a unique minimizer to $\min_{z\in Y} F^k(d,z)$, namely, $\bar{z} = \text{argmax}_{z\in Y} ||z - z^*||$, where $\bar{\tau} = ||\bar{z} - z^*||$. This is no longer true in Case 2. Indeed, the condition of Case 2 implies that $\phi^k(\bar{\tau}; d^k) = 0$ for k sufficiently large, and so $\tau^k = 0$ is an alternate minimizer of $\phi^k(\tau; d^k)$. Moreover, since we have assumed that $z_\infty \notin S(x^*)$ (and hence $z^k \notin S(x^*)$ for k sufficiently large) this implies that $kA^k(z^k)^4 = 0$ and so $kA^k(z^*)^4 = 0$ since $z^k = z^*$ when corresponding to the alternate minimizer to ϕ^k, $\tau^k = 0$. In other words, $\bar{z}$ and z^* are alternate minimizers for the optimization problem in (55). We can recover the desired property of the value function of (55) as a function of d by the following reasoning. By the Theorem of Maximum, $Z^k(d) := \arg\min_{z\in Y} F^k(d,z)$ is upper hemicontinuous in d for k sufficiently large. However, as argued above, we also know that $Z^k(d^k) = Z := \{\bar{z}, z^*\}$ for k sufficiently large. The final step leverages the following result.

lemma:path-connected-continuous

Lemma 6 *In the problem $\min_z h(z|a)$ where h is continuously differentiable with respect to the single-dimensional parameter a, let $z^*(a)$ denote the single-dimensional argmin set as a function of a. If $z^*(a)$ admits a selection that is continuous in a connected neighborhood N of a^* then the value function $H(a) = \min_z h(z|a)$ is continuous and differentiable in the neighborhood N of a^*.*

Proof Let $\tilde{z}^*(a)$ be the selection described in the statement. Take any $a, a' \in N$. Then $H(a) - H(a') = \min_z h(z|a) - \min_z h(z|a') \le h(\tilde{z}^*(a')|a) - h(\tilde{z}^*(a')|a')$, since $\min_z h(z|a) \le h(\tilde{z}^*(a')|a)$ and $\min_z h(z|a') = h(\tilde{z}^*(a')|a')$ since $\tilde{z}^*(a') \in z^*(a)$. Similarly, $H(a) - H(a') \ge h(\tilde{z}^*(a)|a) - h(\tilde{z}(a)|a')$. Therefore, by continuity of $\tilde{z}^*(a)$, $\tilde{z}^*(a') \to \tilde{z}^*(a)$ for $a - a' \to 0^+$ and $\frac{H(a)-H(a')}{a-a'} \le \frac{h(\tilde{z}^*(a')|a)-h(\tilde{z}^*(a')|a')}{a-a'} \to \frac{\partial}{\partial a} h(\tilde{z}^*(a')|a) \to \frac{\partial}{\partial a} h(\tilde{z}^*(a)|a)$, since h is continuously differentiable with respect to a. Similarly, taking $a - a' \to 0^-$ yields $\frac{H(a)-H(a')}{a-a'} \ge \frac{h(\tilde{z}^*(a)|a)-h(\tilde{z}^*(a)|a')}{a-a'} \to \frac{\partial}{\partial a} h(\tilde{z}^*(a)|a)$, again using the continuous differentiability of h with respect to a. Therefore, $H'(a)$ exists and is equal to $\frac{\partial}{\partial a} h(\tilde{z}^*(a)|a)$ in the neighborhood N of a^*. □

We leverage a multi-dimensional implication of this lemma to conclude the directional differentiability of $\min_{z\in Y} F^k(d,z)$ as a function of d in a neighborhood of d^k for k sufficiently large. Moreover, for k sufficiently large (possibly larger) we know that d^* lies in a neighborhood of all d^k (since $d^k \to d^*$) and $Z^k(d^k) = \{\bar{z}, z^*\}$ for k sufficiently large. For clarity, the mapping of the notation in Lemma 6 to our setting is as follows. The parameter a corresponds to d, d^k corresponds to a^*, $z^*(a)$ corresponds to $Z^k(d)$, and $H(a)$ corresponds to $\min_{z\in Y} F^k(d,z)$. The argmin function $Z^k(d)$ satisfies the conditions of the lemma for d sufficiently close to d^k since it will consist of both values $\bar{z}$ and z^* so that a selector continuous in that reason can be chosen.

A.4 Proof of Claim 5

By the Exactness Lemma, $kA^k(z^k)^4 = O(k^{3/4}||z^k - z^*||^2)$. Indeed, if $kA^k(z^k)^4 = \omega(k^{3/4}||z^k - z^*||^2)$ then term (viii) dominates term (vi), and since term (vii) is negative, this only reinforces that $F^k(x^k, y^k, z^k) \to -\infty$ as $k \to \infty$. Hence, it only remains to argue that $kA^k(z^k)^4$ is not $\Theta(k^{3/4}||z^k - z^*||^2)$. By (17), $||z^k - z^*||$ is of constant order, so this amounts to arguing that $kA^k(z^k)^4$ is not $\Theta(k^{3/4})$, or $A^k(z^k)$ is not $\Theta(k^{-1/16})$.

Suppose otherwise that $A^k(z^k) = \Theta(k^{-1/16})$. If the minimum defining $A^k(z^k)$ evaluates to 0 for k sufficiently large, then term $A^k(z^k)$ equals 0, which violates the assumption that $A^k(z^k) = \Theta(k^{-1/16})$. Then clearly, we must assume that $A^k(z^k) < 0$ for all but finitely-many k and so from (18), $A^k(z^k) = P^k + B^k(z^k)$ for all but finitely many k. Observe, in this setting, that

eq:a-b-same-order

$$A^k(z^k) = O(B^k(z^k)). \tag{56}$$

Indeed, either P^k is positive or negative. When P^k is positive and $A^k(z^k) < 0$, it must be that $B^k(z^k)$ dominates and so $A^k(z^k) = O(B^k(z^k))$ holds. When P^k is negative, term (ix) of the penalty function is precisely $-\frac{k}{2}(P^k)^2$. By the Exactness Lemma, which implies that term (ix) converges to 0 as $k \to \infty$, P^k is $o(k^{-1/2})$. Since $A^k(z^k) = \Theta(k^{-1/16})$ it follows that $B^k(z^k)$ must be $\Theta(k^{-1/16})$ and so (56) holds.

Since $A^k(z^k) = O(B^k(z^k))$ (from (56)) and (as we have assumed) $A^k(z^k) = \Theta(k^{-1/18})$, it must be $B^k(z^k) = \Omega(k^{-1/16})$. We now derive a contradiction by showing that, in fact, $B^k(z^k) = o(k^{-1/16})$.

claim:relate-z-to-x

Subclaim 2 $B^k(z^k) = o(k^{-1/16})$.

To see this, first observe that since $\frac{k}{4}\sum_{\ell=1}^p (g_\ell^-(x^k,z^k))^4 \to 0$ (by the Exactness Lemma applying to term (ii) of the penalty function), we have for k sufficiently large

$$\begin{aligned}
-B^k(z^k) &= -f(x^k,z^*) + [f(x^k,z^k) - \tfrac{k}{4}\sum_{\ell=1}^p (g_\ell^-(x^k,z^k))^4 - \tfrac{\alpha_3}{2}\Delta(k)\|z^k-z^*\|^2] \\
&\le -f(x^k,z^*) + [f(x^k,z^k) - \tfrac{k}{4}\sum_{\ell=1}^p (g_\ell^-(x^k,z^k))^{10} - \tfrac{\alpha_3}{2}\Delta(k)\|z^k-z^*\|^2] \\
&= O(\|x^k-x^*\|) - f(x^*,z^*) + f(x^*,z^k) - \tfrac{k}{4}\sum_{\ell=1}^p (g_\ell^-(x^k,z^k))^{10} - \tfrac{\alpha_3}{2}\Delta(k)\|z^k-z^*\|^2 \qquad \text{eq:where-to-continue-from} \quad (57)
\end{aligned}$$

taking the Taylor expansion of x^k around x^* in the difference $f(x^k,z^*) - f(x^k,z^k)$. Note that $k^{1/4}\|d^k - d^*\|^2 = O(1)$ by the exactness and supposition that $A^k(z^k) = \Theta(k^{-1/16})$ (otherwise, term (vii) dominates and $\min_{z\in Y} F^k(x^k,y^k,z) \to -\infty$, a contradiction) and so $O(||x^k - x^*||)$ is $O(k^{-1/8})$.

Next, we derive a lower bound of the term $k\sum_{\ell=1}^p (g_\ell^-(x^k,z^k))^{10}$ in (57) based on the following key inequality:

$$-\frac{k}{4}\sum_{\ell=1}^p g_\ell^-(x^k,z^k)^{10} \le -2^{-9}\frac{k}{4}\sum_{\ell=1}^p g_\ell^-(x^*,z^k)^{10} + Ck^{-1/4} \qquad \text{eq:thing-to-show} \quad (58)$$

for some $C>0$, established at the end of this proof. Using (58) and continuing from (57), we have

$$\begin{aligned}
-B^k(z^k) &\le O(\|x^k-x^*\|) + O(k^{-1/4}) - f(x^*,z^*) + f(x^*,z^k) - \tfrac{k}{4}2^{-9}\sum_{\ell=1}^p (g_\ell^-(x^*,z^k))^{10} - \tfrac{\alpha_3}{2}\Delta(k)\|z^k-z^*\|^2 \\
&\le O(\|x^k-x^*\|) + O(k^{-1/4}) - f(x^*,z^*) + \max_z\{f(x^*,z) - \tfrac{k}{4}2^{-9}\sum_{\ell=1}^p (g_\ell^-(x^*,z))^{10} - \tfrac{\alpha_3}{2}\Delta(k)\|z-z^*\|^2\}. \\
& \qquad \text{eq:this-last-bit} \quad (59)
\end{aligned}$$

By a similar argument as we did in the proof of Claim 2 from (44) onwards we can conclude that , we have $\max_z\{f(x^*,z) - \frac{k}{4}2^{-9}\sum_{\ell=1}^p (g_\ell^-(x^*,z))^{10} - \frac{\alpha_3}{2}\Delta(k)\|z-z^*\|^2\} - f(x^*,z^*)$ is $O(k^{-1/9})$ (complete details are omitted).

Therefore, continuing from (59), we have $-B^k(z^k) \le O(\|x^k-x^*\|) + O(k^{-1/4}) + O(k^{-1/9}) = O(k^{-1/8}) + O(k^{-1/4}) + O(k^{-1/9}) = o(k^{-1/16})$, and thus $B^k(z^k)$ is $o(k^{-1/16})$ as desired for Subclaim 2.

It only remains to establish inequality (58). First, if for some sufficiently large k such that $g_\ell(x^*,z^k) \ge 0$, we have $g_\ell^-(x^*,z^k) = 0$, and inequality (58) is trivially true. Now suppose $g_\ell(x^*,z^k) < 0$. We consider two cases.

Case 1. $g_\ell(x^k,z^k) \le g_\ell(x^*,z^k) < 0$.

This is an easy case since we have

$$-k\sum_{\ell=1}^p g_\ell^-(x^k,z^k)^{10} \le -k\sum_{\ell=1}^p g_\ell^-(x^*,z^k)^{10} \le -2^{-9}k\sum_{\ell=1}^p g_\ell^-(x^*,z^k)^{10} + O(k^{-1/4}).$$

Case 2. $g_\ell(x^k,z^k) - g_\ell(x^*,z^k) > 0$ and $g_\ell(x^*,z^k) < 0$.

We discuss two subcases.

Subcase 1. $0 > g_\ell(x^*,z^k) \ge -[g_\ell^-(x^k,z^k) - g_\ell(x^*,z^k)]$.

In this subcase, we have

$$-g_\ell(x^*,z^k) leg_\ell^-(x^k,z^k) - g_\ell(x^*,z^k) = O(x^k - x^*)\cdot\nabla_x g_\ell(x^*,z^k)) = O(\|x^k-x^*\|),$$

and $k\sum_{\ell=1}^p g_\ell^-(x^k,z^k)^{10} = 0$. It follows

$$k\sum_{\ell=1}^p g_\ell^-(x^*,z^k)^{10} = k\sum_{\ell=1}^p g_\ell^-(x^k,z^k)^{10} + O(k\|x^k-x^*\|^{10}) = k\sum_{\ell=1}^p g_\ell^-(x^k,z^k)^{10} + O(k^{-1/4}),$$

which implies (58) where $k\|x^k-x^*\|^{10} = O(k^{-1/4})$ by exactness $k^{1/4}\|x^k-x^*\|^2 = O(1)$.

Subcase 2. $g_\ell(x^*,z^k) < -[g_\ell^-(x^k,z^k) - g_\ell(x^*,z^k)] < 0$.

In this subcase, $g_\ell^-(x^k, z^k) < 0$. Since $g_\ell^-(x^k, z^k) - g_\ell(x^*, z^k) > 0$, we can write

$$-g_\ell(x^*, z^k) = -[g_\ell(x^*, z^k) + g_\ell^-(x^k, z^k) - g_\ell(x^*, z^k)] + g_\ell(x^k, z^k) - g_\ell(x^*, z^k)$$

and thus by Young's inequality[8]

$$\begin{aligned} g_\ell^-(x^*, z^k)^{10} &= (-g_\ell(x^*, z^k))^{10} \le 2^9(-[g_\ell(x^*, z^k) + g_\ell^-(x^k, z^k) - g_\ell(x^*, z^k)])^{10} + (g_\ell(x^k, z^k) - g_\ell(x^*, z^k))^{10} \\ &= 2^9 g_\ell^-(x^k, z^k)^{10} + (g_\ell(x^k, z^k) - g_\ell(x^*, z^k))^{10}. \end{aligned}$$

Therefore, we have

$$\begin{aligned} -k\sum_{\ell=1}^{p} g_\ell^-(x^k, z^k)^{10} &= -k\sum_{\ell=1}^{p}(-[g_\ell(x^*, z^k) + g_\ell(x^k, z^k) - g_\ell(x^*, z^k)])^{10} \\ &\le -2^{-9}k\sum_{\ell=1}^{p} g_\ell^-(x^*, z^k)^{10} - k\sum_{\ell=1}^{p}(g_\ell(x^k, z^k) - g_\ell(x^*, z^k))^{10} \\ &= -2^{-9}k\sum_{\ell=1}^{p} g_\ell^-(x^*, z^k)^{10} + O(k^{-1/4}), \end{aligned}$$

which implies the desired inequality (58).

A.5 Proof of Claim 6

Again let $d^k = (x^k, y^k) - (x^*, y^*)$. Suppose by way of contradiction that for k sufficiently large,

eq:c-0-equation

$$\begin{aligned} &\left(\tfrac{1}{2}\|z^k - z^*\|^2 - \tfrac{k^{1/4}}{4}\|d^k - d^*\|^2\|z^k - z^*\|^4\right) - \min_z\left\{\tfrac{1}{2}\|z - z^*\|^2 - \tfrac{k^{1/4}}{4}\|d^k - d^*\|^2\|z - z^*\|^4\right\} \\ &\ge C_0 > 0. \end{aligned} \tag{60}$$

Then, for k sufficiently large, let $\tilde{z}^k \in \arg\min_z\{\frac{1}{2}\|z - z^*\|^2 - \frac{k^{1/4}}{4}\|d^k - d^*\|^2\|z - z^*\|^4\}$, and so writing out the sum of terms (vi)–(viii) of the penalty function evaluated at $\tilde{z}^k$ we have for sufficiently large k

$$\begin{aligned} &k^{3/4}\left(\tfrac{1}{2}\|\tilde{z}^k - z^*\|^2 - \tfrac{k^{1/4}}{4}\|d^k - d^*\|^2\|\tilde{z}^k - z^*\|^4\right) - \tfrac{k}{4}A^k(\tilde{z}^k)^4 \\ &\le \tfrac{k^{3/4}}{2}\|z^k - z^*\|^2 - \tfrac{k}{4}\|d^k - d^*\|^2\|z^k - z^*\|^4 - C_0 k^{3/4} - \tfrac{k}{4}A^k(\tilde{z}^k)^4 \\ &< \tfrac{k^{3/4}}{2}\|z^k - z^*\|^2 - \tfrac{k}{4}\|d^k - d^*\|^2\|z^k - z^*\|^4 - \tfrac{k}{4}A^k(z^k)^4, \end{aligned}$$

where the first inequality uses (60) and the second inequality drops the negative term $-kA^k(\tilde{z}^k)^4$ and notes that $kA^k(z^k)^4 = o(k^{3/4})$ by Claim 5. This is a contradiction of the definition of z^k.

A.6 Proof of the "in particular" of Theorem 1

ss:in-particular-proof

If suffices to show that penalty term (vi) converges to 0. We will leverage the key fact established in the earlier part of the proof that $||z^k - z^*|| \to 0$. Now we use the similar argument to show $k^{3/4}\|z^k - z^*\|^2 \to 0$. We discuss two cases, based on the limit of $\lim_{k\to\infty}\frac{k^{1/4}}{4}\|(x^k, y^k) - (x^*, y^*)\|^2 \cdot \|z^k - z^*\|^2$, which is bounded below by $\frac{1}{2}$ by exactness (and discussed earlier in the proof in (49)).

Case 1. $\lim_{k\to\infty}\frac{k^{1/4}}{4}\|(x^k, y^k) - (x^*, y^*)\|^2 \cdot \|z^k - z^*\|^2 = \frac{1}{2}$. In this case, we have terms (vi) and (vii) equal to $k^{3/4}\|z^k - z^*\|^2(\frac{1}{2} - \frac{k^{1/4}}{4}\|d^k - d^*\|^2 \cdot \|z^k - z^*\|^2) = o(k^{3/4}\|z^k - z^*\|^2)$, which implies by exactness that

eq:need-this-part

$$k(A^k(z^k))^4 = o(k^{3/4}\|z^k - z^*\|^2). \tag{61}$$

[8] For any positive number a and b, $(a + b)^{10} = \sum_{s=0}^{10}\begin{pmatrix} s \\ 10 \end{pmatrix} a^{10-s}b^s \le 2^9(a^{10} + b^{10})$ where $a^{10-s}b^s \le \frac{10-s}{10}a^{10} + \frac{s}{10}b^{10}$ and we take $a = -g_\ell(x^*, z^k)$ and $b = g_\ell(x^k, z^k) - g_\ell(x^*, z^k)$.

Therefore, Claim 6 follows since the conclusion of Claim 5 holds in this case. In this case, we can obtain a contradiction if $k^{3/4}\|z^k - z^*\|^2 = \Omega(1)$. Indeed, by the Exactness Lemma,

$$-k(A^k(z^k))^4 + \tfrac{k^{3/2}}{2}\|z^k - z^*\|^2 - \tfrac{k}{4}\|(x^k, y^k) - (x^*, y^*)\|^2 \cdot \|z^k - z^*\|^4 \to 0$$

and when dividing by $k^{3/4}\|z^k - z^*\|^2 = \Omega(1)$ and using (61), we have $\frac{1}{2} - \frac{1}{4}k^{1/4}\|(x^k, y^k) - (x^*, y^*)\|^2 \cdot \|z^k - z^*\|^2 \to 0$, which implies $k^{1/4}\|(x^k, y^k) - (x^*, y^*)\|^2\|z^k - z^*\|^2 \to 2$. However, since z^k is interior as k sufficiently large (by Claim 4), under the assumption $k^{3/4}(z_i^k - z_i^*) \to \infty$, Claim 6 implies the first-order condition of the minimization defined in (19) converges to 0 and so $1 - k^{1/4}\|(x^k, y^k) - (x^*, y^*)\|^2\|z^k - z^*\|^2 \to 0$, a contradiction.

Case 2. $\lim_{k\to\infty} \frac{k^{1/4}}{4}\|(x^k, y^k) - (x^*, y^*)\|^2 \cdot \|z^k - z^*\|^2 = C_1 < \frac{1}{2}$.

In this case,

eq:a-k-goes-0-six-goes-zero

$$kA^k(z^k)^4 = \Theta(k^{3/4}\|z^k - z^*\|^2) \tag{62}$$

by the Exactness Lemma. It suffices to establish the following.

claim:last-thing-to-show

Claim 12 $k(A^k(z^k))^4 = o(k^{3/4}\|z^k - z^*\|^2)$.

If true, this provides the same contradiction as in Case 1. The rest of the proof is to establish Claim 12. We crucially utilize the fact $\|z^k - z^*\| \to 0$. We break the argument into several subclaims, established below.

claim:derivative-first-part

Subclaim 3 $\frac{\partial f(x^k, z^k)}{\partial z_i} - k\sum_{\ell=1}^p g_\ell^-(x^k, z^k)^3 \frac{\partial}{\partial z_i} g_\ell(x^k, z^k) \to 0$ *for every* i.

claim:term-a-under-control

Subclaim 4 $A^k(z^k) = o(||z^k - z^*||)$.

claim:on-the-other-hand

Subclaim 5 $A^k(z^k) = O(||x^k - x^*||)$.

Subclaim 3 is needed to prove Subclaims 4 and 5. With Subclaims 4–5 in hand, we can show Claim 12 by observing

$$\begin{aligned} kA^k(z^k)^4 &= k^{1/4}A^k(z^k)^2 \cdot k^{3/4}(A^k(z^k))^2 \\ &= O(k^{1/4}||x^k - x^*||^2) \cdot o(k^{3/4}||z^k - z^*||^2) \\ &= o(k^{3/4}||z^k - z^*||^2), \end{aligned}$$

where the second equality holds by Subclaims 4 and 5 and the third equality follows since $O(k^{1/4}||x^k - x^*||^2)$ is bounded because penalty term (vii) cannot dominate term (vi) by exactness.

It remains to argue that Subclaims 3–5 hold.

Proof (Proof of Subclaim 3) We want to argue that from term (vii) we get

$$\lim_{k\to\infty}[f(x^k, z^k) - \tfrac{k}{4}\sum_{\ell=1}^p (g_\ell^-(x^k, z^k))^4 - \tfrac{\alpha_3\Delta(k)}{2}||z^k - z^*||^2] = \lim_{k\to\infty}\max_z[f(x^k, z) - \tfrac{k}{4}\sum_{\ell=1}^p (g_\ell^-(x^k, z))^4 - \tfrac{\alpha_3\Delta(k)}{2}||z - z^*||^2]. \tag{63}$$

eq:use-when-not-true

If this is the case then Therefore, the first-order condition of the maximization problem in the right-hand side converges to 0 (by continuous differentiability given any k) and so

$$\frac{\partial f(x^k, z^k)}{\partial z_i} - k\sum_{\ell=1}^p g_\ell^-(x^k, z^k)^3 \frac{\partial}{\partial z_i} g_\ell(x^k, z^k) \to 0$$

since, $\Delta(k) \to 0$ and $||z^k - z^*|| \to 0$. This is the desired conclusion.

eq:use-when-not-true

We establish by contradiction. Suppose otherwise that

$$\lim_{k\to\infty}\{[f(x^k, z^k) - \tfrac{k}{4}\sum_{\ell=1}^p (g_\ell^-(x^k, z^k))^4 - \tfrac{\alpha_3\Delta(k)}{2}||z^k - z^*||^2] - \max_z[f(x^k, z) - \tfrac{k}{4}\sum_{\ell=1}^p (g_\ell^-(x^k, z))^4 - \tfrac{\alpha_3\Delta(k)}{2}||z - z^*||^2]\} = -C_2 < 0.$$

Let $\tilde{z}^k \in \arg\max_z[f(x^k, z) - \frac{k}{4}\sum_{\ell=1}^p (g_\ell^-(x^k, z))^4 - \frac{\alpha_3}{2}\Delta(k)||z - z^*||^2]$. In contradiction to the definition of z^k, we now show that $\tilde{z}^k$ can yield a much smaller value for $F^k(x^k, y^k, z)$ than z^k. Indeed, for k sufficiently large,

$$\begin{aligned} B^k(z^k) - B^k(\tilde{z}^k) = \max_z[f(x^k, z) - \tfrac{k}{4}\sum_{\ell=1}^p (g_\ell^-(x^k, z))^4 - \tfrac{\alpha_3\Delta(k)}{2}||z - z^*||^2] \\ - [f(x^k, z^k) - \tfrac{k}{4}\sum_{\ell=1}^p (g_\ell^-(x^k, z^k))^4 - \tfrac{\alpha_3\Delta(k)}{2}||z^k - z^*||^2] \geq \tfrac{1}{2}C_2 > 0, \end{aligned}$$

and so $B^k(z^k) - B^k(\tilde{z}^k) = \Omega(1)$ and accordingly, $A^k(z^k) + A^k(\tilde{z}^k)$ and $A^k(z^k)^2 + A^k(\tilde{z}^k)^2$ are also $\Omega(1)$. Now, concerning term (vii) of the penalty function, observe that

$$\begin{aligned} &kA^k(z^k)^4 - kA^k(\tilde{z}^k)^4 \\ &= k[A^k(z^k) - A^k(\tilde{z}^k)][A^k(z^k) + A^k(\tilde{z}^k)][A^k(z^k)^2 + A^k(\tilde{z}^k)^2] \\ &= k[B^k(z^k) - B^k(\tilde{z}^k)][A^k(z^k) + A^k(\tilde{z}^k)][A^k(z^k)^2 + A^k(\tilde{z}^k)^2] \end{aligned}$$

and therefore, $k[B^k(z^k) - B^k(\tilde{z}^k)][A^k(z^k) + A^k(\tilde{z}^k)][A^k(z^k)^2 + A^k(\tilde{z}^k)^2] = \Omega(k)$. This implies that term (viii) evaluated at $\tilde{z}^k$ is much smaller than term (viii) evaluated at z^k. This dominates the sublinear term (vi) and so clearly $F^k(x^k, y^k, \tilde{z}^k) < F^k(x^k, y^k, z^k)$ for k sufficiently large, a contradiction. □

Proof (Proof of Subclaim 4) This is immediate from Subclaim 3. By the Taylor expansion of $A^k(z^*)$ around z^k and since $\|z^k - z^*\| \to 0$:

$$\begin{aligned} A^k(z^k) - A^k(z^*) &= \left(\nabla_y f(x^k, z^k) - k\sum_{\ell=1}^{p} g_\ell^-(x^k, z^k)^3 \nabla_y g(x^k, z^k) - \alpha_3 \Delta(k)(z^k - z^*)\right) \cdot (z^k - z^*) + h.o.t \\ &= o(\|z^k - z^*\|). \end{aligned}$$

By exactness, term (ix) yields $A^k(z^*) = o(k^{-1/2}) = o(\|z^k - z^*\|)$ (if $||z^k - z^*|| = O(k^{-1/2})$ then we are done), therefore, we have $A^k(z^k) = o(\|z^k - z^*\|)$. This establishes Subclaim 4. □

Proof (Proof of Subclaim 5) Finally, it remains to argue that Subclaim 5 holds. As argued before in the proof of Subclaim 2, we have

$$\begin{aligned} -B^k(z^k) &= -f(x^*, z^*) + f(x^*, z^k) + O(\|x^k - x^*\|) - \tfrac{k}{4}\sum_{\ell=1}^{p} g_\ell^-(x^k, z^k)^4 - \tfrac{\alpha_3 \Delta(k)}{2}||z^k - z^*||^2 \\ &= -f(x^*, z^*) + f(x^*, z^k) + O(\|x^k - x^*\|) - \tfrac{k}{4}\sum_{\ell=1}^{p} g_\ell^-(x^*, z^k)^4 - \tfrac{\alpha_3 \Delta(k)}{2}||z^k - z^*||^2 \\ &\quad - \tfrac{k}{4}\sum_{\ell=1}^{p} g_\ell^-(x^k, z^k)^4 + \tfrac{k}{4}\sum_{\ell=1}^{p} g_\ell^-(x^*, z^k)^4 \\ &= -f(x^*, z^*) + f(x^*, z^k) + O(\|x^k - x^*\|) - \tfrac{k}{4}\sum_{\ell=1}^{p} g_\ell^-(x^*, z^k)^4 - \tfrac{\alpha_3 \Delta(k)}{2}||z^k - z^*||^2 \\ &\quad - k\sum_{\ell=1}^{p} g_\ell^-(x^k, z^k)^3 \nabla_x g_\ell(x^k, z^k) \cdot (x^k - x^*). \end{aligned}$$

where the last step uses a Taylor's expansion. We now make the following observation that

eq:mfcq-bounded

$$k\sum_{\ell=1}^{p} (g_\ell^-(x^k, z^k))^3 = O(1). \tag{64}$$

Suppose to the contrary that $k\sum_{\ell=1}^{p} g_\ell^-(x^k, z^k)^3$ is not $O(1)$. From Subclaim 3, for all i we have

eq:this

$$\tfrac{\partial}{\partial z_i} f(x^k, z^k) - k\sum_{\ell=1}^{p} (g_\ell^-(x^k, z^k))^3 \tfrac{\partial}{\partial z_i} g_\ell(x^k, z^k) \to 0 \tag{65}$$

as $k \to \infty$. Noting that $(x^k, y^k, z^k) \to (x^*, y^*, z^*)$ by Corollary 1 and Theorem 1, we may rewrite (65) using gradients as $\nabla_y f(x^*, z^*) - \sum_{\ell=1}^{p} \lim_{k\to\infty} kg_\ell^-(x^k, z^k)^3 \nabla_y g_\ell(x^*, z^*) = 0$. We divide both sides of the equation by $||kg(x^k, z^k)^3||$ to yield:

$$\frac{\nabla_y f(x^*, z^*)}{\lim_{k\to\infty} ||kg(x^k, z^k)^3||} - \sum_{\ell=1}^{p} \lim_{k\to\infty} \frac{kg_\ell^-(x^k, z^k)^3}{||kg(x^k, z^k)^3||} \nabla_y g_\ell(x^*, z^*) = 0.$$

The first term above goes to 0 as $k \to \infty$ since $\sum_{\ell=1}^{p} kg_\ell^-(x^k, z^k)^3$ is not $O(1)$ we have $||kg(x^k, z^k)^3|| \to \infty$ as $k \to \infty$. This yields $\sum_{\ell=1}^{p} \lim_{k\to\infty} \frac{kg_\ell^-(x^k, z^k)^3}{||kg(x^k, z^k)^3||} \nabla_y g_\ell(x^*, z^*) = 0$. Observe that the coefficients on $\nabla_y g_\ell(x^*, z^*)$

converge to some ν_ℓ as $k \to \infty$ (with not all $\nu_\ell = 0$) yielding $\sum_{\ell=1}^p \nu_\ell g_\ell(x^*, z^*) = 0$. However, this contradicts (Q1). This establishes (64).

With (64) in hand, we have

$$\begin{aligned}
-B^k(z^k) &\le -f(x^*, z^*) + f(x^*, z^k) + O(\|x^k - x^*\|) - \frac{k}{4}\sum_{\ell=1}^p g_\ell^-(x^*, z^k)^4 - \frac{\alpha_3 \Delta(k)}{2}||z^k - z^*||^2 \\
&\le -f(x^*, z^*) + \max_z\{f(x^*, z) - k\sum_{\ell=1}^p g_\ell^-(x^*, z)^4 - \frac{\alpha_3 \Delta(k)}{2}||z - z^*||^2\} + O(\|x^k - x^*\|) \\
&= \rho(k) - f^* + O(\|x^k - x^*\|) \\
&= O(k^{-1/3}) + O(\|x^k - x^*\|).
\end{aligned}$$

where the first equality uses the definition of $\rho(k)$ in (13) and the second equality uses Claim 2. Therefore, using similar logic to that following (56), $A^k(z^k) = O(B^k(z^k) = O(k^{-1/3}) + O(\|x^k - x^*\|)$. We finally note that $k^{1/3} = O(||x^k - x^*||)$. If $\|x^k - x^*\| = O(k^{-1/3})$ then $kA^k(z^k)^4 \to 0$ and we are done by (62). Thus $A^k(z^k) = O(||x^k - x^*||)$, completing the proof of Subclaim 5. □

A.7 Proof of Claim 7

ss:app-second-derivative-diverges

We proceed by examining the second-order derivative of $F^k(x^k, y^k, z)$ with respect to z_i. Observe that for all i

$$\begin{aligned}
\frac{\partial}{\partial z_i} F^k(x^k, y^k, z) = k^{3/4}\,(z_i - z_i^*) - k(z_i - z_i^*)||(x^k, y^k) - (x^*, y^*)||^2||z - z^*||^2 \\
+ kA^k(z)^3\left(\frac{\partial}{\partial z_i} f(x^k, z) - k\sum_{\ell=1}^p (g_\ell^-(x^k, z))^3 \frac{\partial}{\partial z_i} g_\ell(x^k, z) - \alpha_3 \Delta(k)\,(z_i - z_i^*)\right)
\end{aligned}$$

where $A^k(z)$ is defined in (18), and so

$$\begin{aligned}
\frac{\partial^2}{\partial z_i^2} F^k(x^k, y^k, z) &= k^{3/4} - k(2(z_i - z_i^*)^2 + ||z - z^*||^2)||(x^k, y^k) - (x^*, y^*)||^2 \\
&- \underbrace{3kA^k(z)^2}_{(a)} \underbrace{\left(\frac{\partial}{\partial z_i} f(x^k, z) - k\sum_{\ell=1}^p (g_\ell^-(x^k, z))^3 \frac{\partial}{\partial z_i} g_\ell(x^k, z) - \alpha_3 \Delta(k)\,(z_i - z_i^*)\right)^2}_{(b)} \\
&- \underbrace{kA^k(z)^3}_{(c)} \underbrace{\left(\frac{\partial^2}{\partial z_i^2} f(x^k, z) - k\sum_{\ell=1}^p (g_\ell^-(x^k, z))^3 \frac{\partial^2}{\partial z_i^2} g_\ell(x^k, z) - 3k\sum_{\ell=1}^p (g_\ell^-(x^k, z))^2 (\frac{\partial}{\partial z_i} g_\ell(x^k, z))^2 - \alpha_3 \Delta(k)\right)}_{(d)}.
\end{aligned}$$

eq:second-order-derivative (66)

We now show that

eq:stuff-diverges

$$\frac{\partial^2}{\partial z_i^2} F^k(x^k, y^k, z) \to +\infty \text{ as } k \to \infty \text{ for } z \text{ such that } ||z - z^*|| \le k^{-3/8}. \tag{67}$$

Given (67), by the continuity of $F^k(x^k, y^k, z)$ in z, this implies that $F^k(x^k, y^k, z)$ is strictly convex in z for $||z - z^*|| \le k^{-3/8}$ for k sufficiently large, establishing Claim 7.

It remains to establish (67). If the first term in the second derivative (66) (the term $k^{3/4}$) dominates the other terms, then clearly (67) holds. Hence, if we can show the rest of the terms converge to infinity at a rate slower than $k^{3/4}$ then (67) follows.

The first thing to observe is that the second term in (66) converges to infinity at a slower rate that $k^{3/4}$ for any z such that $||z - z^*|| \le k^{-3/8}$, indeed that term becomes of order $O(kk^{-3/4}) = O(k^{1/4}) = o(k^{3/4})$. Next, we argue about terms (a)–(d) in the following two subclaims. The proofs of these subclaims follow a familiar pattern from earlier arguments. Details are omitted.

claim:derivative-to-zero

Subclaim 6 *For k sufficiently large and for any $z \in B_{k^{-3/8}}(z^*)$, then for all i $\frac{\partial}{\partial z_i} f(x^k, z) - k\sum_{\ell=1}^p (g_\ell^-(x^k, z))^3 \frac{\partial}{\partial z_i} g_\ell(x^k, z) - \alpha_3 \Delta(k)\left(z_i - z_i^*\right) = o(1)$. This expression appears in (b) of* (66).

claim:mfcq-gives-nice-bound

Subclaim 7 *For k sufficiently large and for any z such that $||z - z^*|| \le k^{-3/8}$, $k\sum_{\ell=1}^p (g_\ell^-(x^k, z))^3 = O(1)$. This expression appears in (d) of* (66).

Subclaim 4, along with fact that $||z^k - z^*|| = o(k^{-3/8})$ in Theorem 1, implies that $A^k(z^k) = o(k^{-3/8})$ and (a) in (66) is $o(k^{1/4})$. Together with Subclaim 6, the product of (a) and (b) in (66) is $o(k^{1/4})$.

We now turn to examining terms (c) and (d). From Subclaim 4 we know term (c) is $o(k||z - z^*||^3)$, and thus $o(k^{-1/8})$ by Theorem 1. Moreover, term (d) is $O(k^{1/3})$. Indeed, the first and second terms in (d) are bounded above by constants (the latter follows from Subclaim 7). Moreover, the third term is $O(k^{1/3})$, since Subclaim 7 implies $k \sum_{\ell=1}^{p} (g_\ell^-(x^k, z))^2 (\frac{\partial}{\partial z_i} g_\ell(x^k, z))^2$ is $O(k^{1/3})$. Taken together the product of (c) and (d) is $o(k^{5/24})$. Since $o(k^{5/24})$ is clearly of a smaller order than $k^{3/4}$, the first term in (66) dominates and implies that $\frac{\partial^2}{\partial z_i^2} F^k(x^k, y^k, z)$ diverges to $+\infty$ as $k \to \infty$.

A.8 Proof of Claim 9

ss:app-control-term-ix

If $\gamma^k(\frac{\partial f(x^k,y^k)}{\partial x_i} - \frac{\partial f(x^k,z^*)}{\partial x_i} + \sum_{\ell=1}^{p} \eta_{g,\ell}^k \frac{\partial g_\ell(x^k,y^k)}{\partial x_i}) \to 0$ for all i, then we obtain from (28):

$$0 \leftarrow \nabla_x F(x^k, y^k) + \sum_{j=1}^{q} \eta_{G,j}^k \nabla_x G_j(x^k, y^k) + \sum_{\ell=1}^{p} \eta_{g,\ell}^k \nabla_x g_\ell(x^k, y^k) - \sum_{\ell \in A_g(x^*,z^*)} \xi_\ell^k \nabla_x g_\ell(x^k, z^*) + \sum_{i=1}^{m} \vartheta_i^k \nabla_x \frac{\partial f(x^k,y^k)}{\partial y_i}.$$

However, as argued at the outset of the proof, all remaining coefficients are bounded and so dividing by their norm yields an optimality condition of the form (33a)–(33b) with $\mu = 0 \neq 1$. This is a contradiction. Thus, it remains to consider the case

$$\gamma^k ||\nabla_x f(x^k, y^k) - \nabla_x f(x^k, z^*) + \sum_{\ell=1}^{p} \eta_{g,\ell}^k \nabla_x g_\ell(x^k, y^k)|| = \Omega(1).$$

However, the above object is bounded (that is, $O(1)$). This is implied by the first-order condition $\nabla F^k(x^k, y^k, z^k) = 0$ and isolating the term involving γ which yields:

$$\begin{aligned} &\gamma^k \left(\nabla_x f(x^k, y^k) - \nabla_x f(x^k, z^*) + \sum_{\ell=1}^{p} \eta_{g,\ell}^k \nabla_x g_\ell(x^k, y^k) \right) \\ &\to -\Bigg(\nabla_x F(x^k, y^k) + \sum_{j=1}^{q} \eta_{G,j}^k \nabla_x G_j(x^k, y^k) + \sum_{\ell=1}^{p} \eta_{g,\ell}^k \nabla_x g_\ell(x^k, y^k) \\ &\quad - \sum_{\ell \in A_g(x^*,z^*)} \xi_\ell^k \nabla_x g_\ell(x^k, z^*) + \sum_{i=1}^{m} \vartheta_i^k \nabla_x \frac{\partial f(x^k,y^k)}{\partial y_i} \Bigg) = O(1), \end{aligned}$$

where the final equality follows since all of the coefficients are bounded and the gradients are all bounded over the compact feasible region $X \times Y$. This implies that $\gamma^k ||\nabla_x f(x^k, y^k) - \nabla_x f(x^k, z^*) + \sum_{\ell=1}^{p} \eta_{g,\ell}^k \nabla_x g_\ell(x^k, y^k)|| = \Theta(1)$ and establishes the statement of the claim.

A.9 Proof of Claim 10

ss:app-control-gamma

Now using the first order condition $0 = \nabla F^k(x^k, y^k, z^k)$. If we multiply the above first order condition by a vector $u^k = (d_x^k, d_y^k) = (x^k - x^*, y^k - y^*)$, yields $(\nabla F^k(x^k, y^k, z^k))^\top u^k = 0$. Note all terms in the left-hand side of $(\nabla F^k(x^k, y^k, z^k))^\top u^k = 0$ other than term that associated with penalty term (ix) are $O(||u^k||)$ by the assumption that $\gamma^k \to \infty$. Thus,

eq:isolating

$$\begin{aligned} &k \min\{0, f(x^k, y^k) - \tfrac{k}{4} \sum_{\ell \in A_g} g_\ell^-(x^k, y^k)^4 - f(x^k, z^*)\} \times \\ &[\nabla f(x^k, y^k) - k \sum_{\ell \in A_g} g_\ell^-(x^k, y^k)^3 \nabla g_\ell(x^k, y^k) - \nabla f(x^k, z^*)]^\top (u^k) = O(||u^k||). \end{aligned} \tag{68}$$

Moreover, by taking a Taylor expansion around (x^k, y^k), we can write

eq:a-tricky-thing

$$\begin{aligned} &f(x^k, y^k) - \tfrac{k}{4} \sum_{\ell \in A_g} g_\ell^-(x^k, y^k)^4 - f(x^k, z^*) \\ &= [\nabla f(x^k, y^k) - k \sum_{\ell \in A_g} g_\ell^-(x^k, y^k)^3 \nabla g_\ell(x^k, y^k) - \nabla f(x^k, z^*)]^\top (u^k) + o(\|u^k\|) \end{aligned} \tag{69}$$

where the second equality uses the fact that $f(x^*, y^*) = f(x^*, z^*)$ and $g_\ell^-(x^*, y^*) = 0$ since (x^*, y^*) is feasible, and since all the gradients are bounded. Now, observe that

$$\begin{aligned}
O(||u^k||) &= k\min\{0, f(x^k, y^k) - \tfrac{k}{4}\sum_{\ell\in A_g} g_\ell^-(x^k, y^k)^4 - f(x^k, z^*)\}[\nabla f(x^k, y^k) - k\sum_{\ell\in A_g} g_\ell^-(x^k, y^k)^3 \nabla g_\ell(x^k, y^k) - \nabla f(x^k, z^*)]^\top (u^k) \\
&= \Theta(k\min\{0, f(x^k, y^k) - \tfrac{k}{4}\sum_{\ell\in A_g} g_\ell^-(x^k, y^k)^4 - f(x^k, z^*)\}[\min\{0, f(x^k, y^k) - \tfrac{k}{4}\sum_{\ell\in A_g} g_\ell^-(x^k, y^k)^4 - f(x^k, z^*)\} + o(||u^k||)] \\
&= \Theta(k\min\{0, f(x^k, y^k) - \tfrac{k}{4}\sum_{\ell\in A_g} g_\ell^-(x^k, y^k)^4 - f(x^k, z^*)\}^2)
\end{aligned}$$

where the first equality is (68), the second equality is (69) and the third equality uses the fact that term associated with $o(||u^k||)$ is of order smaller than $O(||u^k||)$.

It thus follows that

$$(\gamma^k)^2 = \Big(k\min\{0, f(x^k, y^k) - \tfrac{k}{4}\sum_{\ell=1}^{p} g_\ell^-(x^k, y^k)^4 - f(x^k, z^*)\}^2\Big)^2 = O(k\|u^k\|)$$

or $\gamma^k = O(k^{1/2}\|u^k\|^{1/2})$. This establishes the claim.

A.10 Proof of Claim 11

ss:app-d-k-gets-large

Since $\gamma^k = O(k^{1/2}\|u^k\|^{1/2})$, we consider two cases, either $\gamma^k = o(k^{1/2}\|u^k\|^{1/2})$ or $\gamma^k = \Theta(k^{1/2}\|u^k\|^{1/2})$. We treat the former. From the first-order condition $\nabla F^k(x^k, y^k, z^k) = 0$, we have

gamma-term1

$$\gamma^k \nabla(f(x^k, y^k) - f(x^k, z^*)) = \Theta(1), \tag{70}$$

for any dimension of (x, y) since all coefficients other than γ^k are bounded. Note also that we are in the case $\|\nabla(f(x^*, y^*) - f(x^*, z^*))\| = 0$, it must be for any i, $\frac{\partial f(x^k,y^k)}{\partial x_i} - \frac{\partial f(x^k,z^*)}{\partial x_i} = \nabla(\frac{\partial f(x^*,y^*)}{\partial x_i} - \frac{\partial f(x^*,z^*)}{\partial x_i})^\top u^k + o(\|u^k\|) = O(\|u^k\|)$ where we use the Taylor expansion around (x^*, y^*) and leverage the fact that $\|\nabla(f(x^*, y^*) - f(x^*, z^*))\| = 0$. Therefore, from (70) we have $\gamma^k \nabla(f(x^k, y^k) - f(x^k, z^*)) = o(k^{1/2}\|u^k\|^{1/2}) \cdot O(\|u^k\|) = \Theta(1)$, which implies $k^{1/2}\|u^k\|^{3/2} \to \infty$ or equivalently $k\|u^k\|^3 \to \infty$. That is, $k||u^k||^3 = \omega(1)$. When $\gamma^k = \Theta(k^{1/2}||u^k||^{1/2})$ a similar argument shows $k||u^k||^3 = \Theta(1)$.