Spurious Stationarity and Hardness Results for Mirror Descent

He Chen HCHEN@SE.CUHK.EDU.HK
The Chinese University of Hong Kong

Jiajin Li JIAJIN.LI@SAUDER.UBC.CA
The University of British Columbia

Anthony Man-Cho So MANCHOSO@SE.CUHK.EDU.HK
The Chinese University of Hong Kong
Authors ordered alphabetically

Abstract

Despite the success of Bregman proximal-type algorithms, such as mirror descent, in machine learning, most theoretical results depend on the gradient Lipschitz property of the kernel, excluding widely used cases like the Shannon entropy kernel. This paper uncovers a fundamental limitation: *Spurious stationary points* inevitably arise when non-gradient Lipschitz kernels are used. We establish an algorithm-dependent hardness result, showing that Bregman proximal-type algorithms cannot escape these spurious stationary points in finite steps if the initial point is unfavorable, even in convex settings. Those challenges are discovered through the lack of a well-defined stationarity measure, typically based on Bregman divergence, for these algorithms. While some extensions attempt to address this, we demonstrate that they still fail to distinguish reliably between stationary and non-stationary points for non-gradient Lipschitz kernels. Our findings highlight the need for new theoretical tools and algorithms within Bregman geometry, opening new avenues for further research.

1. Introduction

In this paper, we study structured nonsmooth (non)-convex optimization problems of the form

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} F(\boldsymbol{x}) := f(\boldsymbol{x}) + g(\boldsymbol{x}), \tag{P}$$

where $\mathrm{dom}(g) = \mathcal{X}$ is a nonempty closed convex set, $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function, and $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a convex and locally Lipschitz continuous function. To solve Problem (P) efficiently, Bregman proximal-type (BPs) algorithms are widely used methods that effectively exploit the geometry of the set $\mathcal{X}$ to avoid costly projection or proximal steps under the Ecludiean space [1, 6, 8, 26].

We now present the unified formulation of the class of Bregman proximal-type (BPs) algorithms that we will investigate in this paper, where the iterative schemes are defined as follows

$$\boldsymbol{x}^{k+1} = \operatorname*{argmin}_{\boldsymbol{y}\in\mathbb{R}^n} \left\{ \gamma\left(\boldsymbol{y}; \boldsymbol{x}^k\right) + g(\boldsymbol{y}) + \frac{1}{t_k} D_h(\boldsymbol{y}, \boldsymbol{x}^k) \right\}. \tag{1.1}$$

Here, $\gamma(\,\cdot\,;\boldsymbol{x})$ is the surrogate model for f at the point $\boldsymbol{x}$, $t_k \geq 0$ is the step size in the k-th iteration and D_h is the Bregman divergence associated with the kernel function h. Specifically, when the

surrogate model is the original function f, the update rule (1.1) simplifies to the Bregman proximal point method [11, 17]. If the surrogate model is a linear approximation, given by $\gamma(\boldsymbol{y}; \boldsymbol{x}^k) := f(\boldsymbol{x}^k) + \nabla f(\boldsymbol{x}^k)^T(\boldsymbol{x} - \boldsymbol{x}^k)$, then the update (1.1) encompasses the Bregman proximal gradient descent method (BPG) [3, 5, 10, 24, 29]. Alternatively, one may choose a quadratic surrogate model given by $\gamma(\boldsymbol{y}; \boldsymbol{x}^k) := f(\boldsymbol{x}^k) + \nabla f(\boldsymbol{x}^k)^T(\boldsymbol{x} - \boldsymbol{x}^k) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^k)^T\nabla^2 f(\boldsymbol{x}^k)(\boldsymbol{x} - \boldsymbol{x}^k)$, which has been recently investigated by Doikov and Nesterov [13].

Many efforts have been devoted to obtaining non-asymptotic convergence results for Bregman proximal-type algorithms, which are expected to be similar to those for Euclidean cases. However, these results have been limited to scenarios where $\mathrm{dom}(h) = \mathbb{R}^n$ (see, e.g., [27]) or where ∇h is Lipschitz continuous (see, e.g., [28]). These conditions essentially guarantee the non-degenerate property of the mirror map, where Euclidean and Bregman geometries align. Unfortunately, they exclude most of powerful kernels that enhance the practical appeal of BPs, such as the Shannon entropy function. This raises a natural question: *Is there any fundamental difference between Bregman proximal-type algorithms and their Euclidean counterparts when the mirror map is degenerate (i.e., when the kernel function is not gradient Lipschitz)?*

In this paper, we fully address this question and explore the fundamental differences that prevent achieving analogous convergence results for non-gradient Lipchitz kernels. We begin by identifying a class of undesirable points, termed *spurious stationary points*, which arise in scenarios where the kernel function is not gradient Lipschitz continuous. We then show that the existence of spurious stationary points is inevitable, even in simple convex problems. More importantly, if we initialize BPs near a spurious stationary point, the generated sequence can remain trapped within a small neighborhood of this point, regardless of the number of steps taken. This hardness result provides a negative answer to the possibility of obtaining similar convergence results as those in the Euclidean space.

Taking a step further, we aim to clarify why BPs become trapped at these spurious stationary points and how we discovered this result that has been overlooked by the research community for an extended period. To analyze the behavior of the iterate sequences and their proximity to stationary points, researchers typically introduce a residual function $R : \mathbb{R}^n \to \mathbb{R}_+$ quantifies the stationarity of the iterates and subsequently establish the convergence of $\{R(\boldsymbol{x}^k)\}_{k\in\mathbb{N}}$. This trapping phenomenon is observed through the unsatisfactory behaviours of all existing stationarity measures. Specifically, these measures fail to satisfy the following equivalence (Q) when the kernel function is not gradient Lipschitz continuous:

$$\boxed{\lim_{k\to\infty} R(\boldsymbol{x}^k) = 0 \underset{?}{\iff} \mathbf{0} \in \partial F\left(\lim_{k\to\infty} \boldsymbol{x}^k\right).} \tag{Q}$$

This failure implies that we cannot classify whether the iterates of BPs are approximately stationary or not, even when the stationarity measures are relatively small. Spurious stationary points are precisely those points where the stationarity measure equals zero, yet zero does not belong to their subdifferential. It is worth noting that the equivalence (Q) trivially holds for all algorithms under Euclidean geometry, as well as under conditions such as $\mathrm{dom}(h) = \mathbb{R}^n$ (see, e.g., [27]) or when ∇h is Lipschitz continuous (see, e.g., [28]) for BPs. Consequently, there are no spurious stationary points in these cases.

Main techinical novelty The key tool for establishing the existence of spurious stationary points and the associated hardness results is the *extended Bregman stationarity measure*, which may also

be of independent interest. We begin by unifying all existing stationarity measures and extending their definitions to encompass the entire domain of kernels. This extension allows us to compute the residual function at accumulation points. Notably, prior to our work, all existing stationarity measures were not well-defined at the boundaries of the kernel functions' domains. We then establish the continuity properties of the newly developed stationarity measure and demonstrate that a stationarity measure equal to zero can still indicate non-stationary points (i.e., spurious stationary points). The hardness results follow directly from this analysis.

Notation. Our notation is mostly standard. The set of extended real numbers is denoted by $\overline{\mathbb{R}}$. For a vector $\boldsymbol{x} \in \mathbb{R}^n$, its i-th coordinate is represented by x_i, and $\boldsymbol{x}_{\mathcal{I}}$ denotes a subvector of $\boldsymbol{x}$ indexed by $\mathcal{I}$. The Euclidean ball $\mathbb{B}_\epsilon(\boldsymbol{x})$ is defined through $\mathbb{B}_\epsilon(\boldsymbol{x}) := \{\boldsymbol{y} \in \mathbb{R}^n : \|\boldsymbol{x} - \boldsymbol{y}\|_2 \leq \epsilon\}$. Given a set $\mathcal{X} \subseteq \mathbb{R}^n$, we use $\mathrm{cl}(\mathcal{X})$, $\mathrm{int}(\mathcal{X})$, and $\mathrm{bd}(\mathcal{X})$ to denote its closure, interior, and boundary, respectively. The indicator function $\delta_{\mathcal{X}}$ of the set $\mathcal{X}$ is defined through $\delta_{\mathcal{X}}(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \in \mathcal{X}$; $\delta_{\mathcal{X}}(\boldsymbol{x}) = +\infty$ otherwise. We employ the shorthand $\delta_{\mathbf{g}(\boldsymbol{x})=\mathbf{0}}$ to compactly represent $\delta_{\{\boldsymbol{x}\in\mathbb{R}^n:\mathbf{g}(\boldsymbol{x})=\mathbf{0}\}}$ for any real-valued function $g : \mathbb{R}^n \to \mathbb{R}$.

2. Assumptions and justification

In this section, we state our blanket assumptions and justify their validity in applications. To begin, we introduce the definition of the separable kernel function.

Definition 1 *We say that $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a separable kernel function if (i) h is defined by $h(\boldsymbol{x}) = \sum_{i=1}^n \varphi(x_i)$, where $\varphi : \mathbb{R} \to \overline{\mathbb{R}}$ is a univariate function that is continuously differentiable on $\mathrm{int}(\mathrm{dom}(\varphi))$; (ii) $|\varphi'(x^k)| \to +\infty$ as $x^k \to x \in \mathrm{bd}(\mathrm{dom}(\varphi))$; (iii) φ is strictly convex.*

The separability structure outlined in property (i) is ubiquitous in real-world scenarios [2, 5, 20]. Properties (ii) and (iii) are referred to as Legendre-type properties, as defined in [3]. We illustrate their practicality by presenting several widely used kernel functions in Example 1.

Assumption 1 (Problem (P)) *Let $\mathrm{dom}(g) = \mathcal{X}$ be a nonempty closed convex set. Suppose:*

(i) *The function f is continuously differentiable on $\mathcal{X}$.*

(ii) *The function g is convex and locally Lipschitz continuous on $\mathcal{X}$.*

(iii) *There exists a strictly feasible point $\boldsymbol{x}^{\mathrm{int}} \in \mathrm{int}(\mathrm{dom}(h)) \cap \mathcal{X}$ and $\mathcal{X} \subseteq \mathrm{cl}(\mathrm{dom}(h))$.*

(iv) *The function h is a separable kernel function, see Definition 1.*

Assumptions 1 (i)-(iii) or their stronger versions are widely adopted in the literature; see, e.g., Assumption A in [3], Assumption A in [5], and Definition 1 and Assumption 1 in [2]. We also refer the readers to see the practical problems that satisfy these assumptions and are solved by Bregman proximal-type methods satisfying these assumptions in, e.g., problem (3) in [23], and problem (6) in [24] and problem (7) in [12].

Next, we proceed to give assumptions of algorithm classes (i.e., the update rule (1.1)), which is characterized by the surrogate model γ.

Assumption 2 (Surrogate model γ) *The following hold.*

(i) *The function $(\boldsymbol{x}, \boldsymbol{y}) \mapsto \gamma(\boldsymbol{y}; \boldsymbol{x})$ and gradient $(\boldsymbol{x}, \boldsymbol{y}) \mapsto \nabla\gamma(\boldsymbol{y}; \boldsymbol{x})$ (w.r.t $\boldsymbol{y}$) are jointly continuous w.r.t. $(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{y} \in \mathcal{X}$ and $\boldsymbol{x} \in \mathcal{X}$.*

(ii) *For all* $\boldsymbol{x} \in \mathcal{X}$, *we have* $\nabla\gamma(\boldsymbol{y};\boldsymbol{x})\,|_{\boldsymbol{y}=\boldsymbol{x}} = \nabla f(\boldsymbol{x})$, *and* $\gamma(\boldsymbol{y};\boldsymbol{x})\,|_{\boldsymbol{y}=\boldsymbol{x}} = f(\boldsymbol{x})$.

(iii) *For all* $\boldsymbol{x} \in \mathcal{X}$, *there exists a constant* $\bar{t} > 0$ *such that* $\bar{t}\gamma(\,\cdot\,;\boldsymbol{x}) + h(\cdot)$ *is strictly convex.*

(iv) *Either* $\mathcal{X}$ *is compact or we have the following condition: For all step sizes* $t \in (0, \bar{t}]$ *and all sequences* $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}, \{\boldsymbol{y}^k\}_{k\in\mathbb{N}} \subseteq \operatorname{int}(\operatorname{dom}(h)) \cap \mathcal{X}$ *such that* $\|\boldsymbol{y}^k\| \to \infty$ *and* $\boldsymbol{x}^k \to \overline{\boldsymbol{x}} \in \mathcal{X}$, *the following holds:*

$$\lim_{k\to\infty} \gamma(\boldsymbol{y}^k;\boldsymbol{x}^k) + g(\boldsymbol{y}^k) + \frac{1}{t}D_h(\boldsymbol{y}^k,\boldsymbol{x}^k) = +\infty, \tag{2.1}$$

where $D_h(\boldsymbol{y},\boldsymbol{x}) := h(\boldsymbol{y}) - h(\boldsymbol{x}) - h'(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})$.

Unless otherwise specified, the step size t *in this paper is assumed to satisfy* $t \in (0, \bar{t}]$.

Assumptions 2 (i) and (ii) are standard, serving to ensure the continuity and local correctness of the surrogate model γ. Assumption 2 (iii) usually reduces to conditions commonly adopted in the literature or is automatically satisfied for all three choices mentioned in the introduction. When γ represents the first-order expansion of f at the current iterate $\boldsymbol{x}$, Assumption 2 (iii) is trivially satisfied. If γ is selected as the original function f, it reduces to the relatively weak convexity condition, as discussed in [9, 28]. When γ is the second-order expansion of f at the current iterate $\boldsymbol{x}$, the L-smoothness of f and strong convexity of h suffice to guarantee Assumption 2 (iii). Assumption 2 (iv) is to ensure the well-posedness of the Bregman proximal-type update (1.1). If $\mathcal{X}$ is unbounded, then Assumption 2 (iv) can be implied by supercoercive-type conditions, see Lemma 2 in Bauschke et al. [3] and Assumption B in Bolte et al. [9]. Interested readers are referred to Appendix B for the rigorous verification of this assumption for commonly used γ.

Based on Assumptions 1 and 2, we present the following lemma, which discusses the well-posedness of update (1.1) on $\operatorname{int}(\operatorname{dom}(h))$. Similar results can also be found in [3, 14].

Lemma 2 *For all* $\boldsymbol{x} \in \operatorname{int}(\operatorname{dom}(h)) \cap \mathcal{X}$, *we have* $T_\gamma^t(\boldsymbol{x}) \in \operatorname{int}(\operatorname{dom}(h)) \cap \mathcal{X}$, *where the update mapping* T_γ^t *is defined through* $T_\gamma^t := \operatorname{argmin}_{\boldsymbol{y}}\ \gamma(\boldsymbol{y};\boldsymbol{x}^k) + g(\boldsymbol{y}) + \frac{1}{t}D_h(\boldsymbol{y},\boldsymbol{x}^k)$.

Algorithm 1: Bregman divergence-based algorithms

Input: $\{t_k\}_{k\in\mathbb{N}}$ with $t_k > 0$, initial point $\boldsymbol{x}^0 \in \operatorname{int}(\operatorname{dom}(h)) \cap \mathcal{X}$, and $k = 0$.
while *Stopping Criteria is not satisfied* **do**
  $\boldsymbol{x}^{k+1} = T_\gamma^{t_k}(\boldsymbol{x}^k)$;
  Set $k = k + 1$
end
Output: The last iterate $\boldsymbol{x}^k$

3. Main results

In this section, we aim to provide a negative answer for (Q) and further establish a hardness result for the BPs. The central tool to achieve these results is a newly introduced notion, namely, *spurious stationary points*. To begin, we define the index sets

$$\mathcal{B}(\boldsymbol{x}) := \{b \in [n] : x_b \in \operatorname{bd}(\operatorname{dom}(\varphi))\}; \quad \mathcal{I}(\boldsymbol{x}) := \{i \in [n] : x_i \in \operatorname{int}(\operatorname{dom}(\varphi))\}.$$

Then, the formal definition of spurious stationary points is given as follows.

Definition 3 (Spurious stationary points) *A point $\boldsymbol{x} \in \mathcal{X}$ is defined as a spurious stationary point of problem (P) if there exists a vector $\boldsymbol{p} \in \partial F(\boldsymbol{x})$ satisfying $\boldsymbol{p}_{\mathcal{I}(\boldsymbol{x})} = 0$ but $\mathbf{0} \notin \partial F(\boldsymbol{x})$.*

Remark 4 *(i) Spurious stationary points exist only when the kernel is non-gradient Lipschitz. (ii) For a kernel h with gradient Lipschitz property, Definition 1 (ii) implies that $\mathrm{dom}(h) = \mathbb{R}^n$ and $\mathcal{I}(\boldsymbol{x}) = [n]$ hold for all $\boldsymbol{x} \in \mathcal{X}$, thereby precluding the existence of spurious stationary points by definition.*

The following proposition demonstrates that spurious stationary points are precisely counter examples where zero does not belong to the subdifferential, yet the stationary measure equals zero. Consequently, one *cannot* classify whether the output points of the BPs are approximately stationary even when their stationarity measure is small.

Proposition 5 *Let $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}} \subseteq \mathrm{int}(h) \cap \mathcal{X}$ be a sequence converging to a spurious stationary point $\tilde{\boldsymbol{x}}^* \in \mathcal{X}$ for problem (P). Then, the Bregman stationarity measure of $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$ goes to zero, i.e., for all $t > 0$ and $R(\boldsymbol{x}) := D_h(T_\gamma^t(\boldsymbol{x}), \boldsymbol{x})$ or $R(\boldsymbol{x}) := D_h\left(\mathrm{prox}_{t,\gamma}^h(\boldsymbol{x}), \boldsymbol{x}\right)$,*

$$\lim_{k\to\infty} R(\boldsymbol{x}^k) = 0.$$

Remark 6 *We consider $R(\boldsymbol{x}) := D_h(T_\gamma^t(\boldsymbol{x}), \boldsymbol{x})$ and $R(\boldsymbol{x}) := D_h\left(\mathrm{prox}_{t,\gamma}^h(\boldsymbol{x}), \boldsymbol{x}\right)$ in Proposition 5 because these two measures typically evaluate the descent property of the BPs; see e.g., [3, 28]. Other measures, such as the function value gap $F(\boldsymbol{x}) - F^*$ and the minimal subgradient norm $\mathrm{dist}(\mathbf{0}, \partial F(\boldsymbol{x}))$, either fail to identify stationary points in the nonconvex setting or have no decrease guarantee for the iterate sequence of the BPs, as illustrated by Example 3.*

We further uncover some practical challenges stemming from spurious stationary points. While it seems that spurious stationary points only impact the Bregman stationarity measure and that the BPs will get rid of them since they are non-stationary, the following hardness result demonstrates that we cannot escape from spurious stationary points in finite steps using Algorithm 1.

Theorem 7 (Hardness) *If there exists a spurious stationary point $\tilde{\boldsymbol{x}}^* \in \mathcal{X}$ for problem (P), then for every $K \in \mathbb{N}$ and $\epsilon > 0$, there exists an initial point $\boldsymbol{x}^0 \in \mathbb{B}_\epsilon(\tilde{\boldsymbol{x}}^*) \cap \mathcal{X} \cap \mathrm{int}(h)$, sufficiently close to the spurious point $\tilde{\boldsymbol{x}}^*$, such that the sequence $\{\boldsymbol{x}_k\}_{k\in[K]}$ generated by Algorithm 1 satisfies*

$$\boldsymbol{x}^k \in \mathbb{B}_\epsilon(\tilde{\boldsymbol{x}}^*) \quad \text{for all } k \in [K]. \tag{3.1}$$

Remark 8 (i) *If the iterates of Algorithm 1 enter a small neighborhood of a spurious stationary point, then we cannot get rid of them with any finite steps.* (ii) *Allowing for an infinite number of steps, under certain conditions, Algorithm 1 can eventually escape from them and converge to true stationary points. For instance, Corollary 1 in [3] establishes that, under the convexity of f and additional conditions, the sequence $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$ generated by BPG satisfies*

$$f(\boldsymbol{x}^k) - \min_{\boldsymbol{x}\in\mathcal{X}} f \leq \frac{D_h(\overline{\boldsymbol{x}}, \boldsymbol{x}^0)}{t} \cdot \frac{1}{k}, \tag{3.2}$$

where $\overline{\boldsymbol{x}} \in \mathrm{argmin}_{\boldsymbol{x}\in\mathcal{X}} f$ is the global minimizer, t is the step size, and $\boldsymbol{x}^0$ is an arbitrary initial point. The non-asymptotic convergence result (3.2) guarantees that $f(\boldsymbol{x}^k) \to \min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$.

Therefore, the limit points of $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$ *are globally optimal (stationary).* (iii) *Our hardness result does not contradict the non-asymptotic rate in* (3.2)*. When* (3.1) *holds, the initial point* $\boldsymbol{x}^0$ *is sufficiently close (as we constructed) to a spurious stationary point* $\tilde{\boldsymbol{x}}^*$*, and the distance* $D_h(\overline{\boldsymbol{x}}, \boldsymbol{x}^0)$ *can be extremely large, allowing* (3.2) *to hold for all* $k \in [K]$.

A remaining question is that whether spurious stationary points exist in practical optimization problems. In Appendix F, we give examples of spurious stationary points in both convex and nonconvex problems. Furthermore, the following proposition demonstrates that for a broad class of convex problems with polytopal constraints, the existence of spurious points is guaranteed.

Proposition 9 (Existence of spurious stationary points) *Consider a convex optimization problem* $\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$*, where* $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^n : A\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}\}$ *is compact,* $A \in \mathbb{R}^{m\times n}$*, and* f *is not constant on* $\mathcal{X}$*. If Assumption 1 holds and* $\mathrm{cl}(\mathrm{dom}(h)) = \mathbb{R}^n_+$*, then every maximal point* $\tilde{\boldsymbol{x}}^* \in \mathrm{argmax}_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$ *is a spurious stationary point.*

4. Closing remark

The algorithm-dependent hardness results in this paper raise an open question:

How can we escape spurious stationary points by modifying Bregman proximal-type algorithms and establish a valid non-asymptotic convergence rate?

To address this question, our paper suggests one possible direction: Identify a stationarity measure that meets the equivalence (Q) and has a decreasing guarantee for the BP iterates. For convex settings, a well-known candidate stationarity measure is the function value gap $F - F^*$. However, when the objective functions are nonconvex, the possibility of addressing this question remains unanswered. By seeking an eligible stationarity measure, one might discover new principles for algorithm design to eliminate undesirable points.

References

[1] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of computing*, 8(1):121–164, 2012.

[2] Waiss Azizian, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the rate of convergence of Bregman proximal methods in constrained variational inequalities. *arXiv preprint arXiv:2211.08043*, 2022.

[3] Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

[4] Heinz H. Bauschke, Minh N. Dao, and Scott B. Lindstrom. Regularizing with Bregman–Moreau envelopes. *SIAM Journal on Optimization*, 28(4):3208–3228, 2018.

[5] Heinz H. Bauschke, Jérôme Bolte, Jiawei Chen, Marc Teboulle, and Xianfu Wang. On linear convergence of non-Euclidean gradient methods without strong convexity and Lipschitz gradient continuity. *Journal of Optimization Theory and Applications*, 182(3):1068–1087, 2019.

[6] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[7] Amrit Singh Bedi, Souradip Chakraborty, Anjaly Parayil, Brian M. Sadler, Pratap Tokekar, and Alec Koppel. On the hidden biases of policy mirror ascent in continuous action spaces. In *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*, pages 1716–1731. PMLR, 2022.

[8] Benjamin Birnbaum, Nikhil R Devanur, and Lin Xiao. Distributed algorithms via gradient descent for Fisher markets. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 127–136, 2011.

[9] Jérôme Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.

[10] Yair Censor and Stavros Andrea Zenios. Proximal minimization algorithm with D-functions. *Journal of Optimization Theory and Applications*, 73(3):451–464, 1992.

[11] Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

[12] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, pages 1542–1553. PMLR, 2020.

[13] Nikita Doikov and Yurii Nesterov. Gradient regularization of newton method with Bregman distances. *Mathematical Programming*, 204(1):1–25, 2023.

[14] Radu-Alexandru Dragomir, Adrien B Taylor, Alexandre d'Aspremont, and Jérôme Bolte. Optimal complexity and certification of Bregman first-order methods. *Mathematical Programming*, 194(1):41–83, 2022.

[15] Feihu Huang, Shangqian Gao, and Heng Huang. Bregman gradient policy optimization. In *Proceedings of the 10th International Conference on Learning Representations (ICLR 2022)*, 2022.

[16] Feihu Huang, Junyi Li, Shangqian Gao, and Heng Huang. Enhanced bilevel optimization via Bregman distance. In *Advances in Neural Information Processing Systems 35*, pages 28928–28939, 2022.

[17] Krzysztof C. Kiwiel. Proximal minimization methods with generalized Bregman functions. *SIAM journal on Control and Optimization*, 35(4):1142–1168, 1997.

[18] Puya Latafat, Andreas Themelis, Masoud Ahookhosh, and Panagiotis Patrinos. Bregman Finito/MISO for nonconvex regularized finite sum minimization without Lipschitz gradient continuity. *SIAM Journal on Optimization*, 32(3):2230–2262, 2022.

[19] Tim Tsz-Kit Lau and Han Liu. Bregman proximal Langevin Monte Carlo via Bregman-Moreau Envelopes. In *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*, pages 12049–12077. PMLR, 2022.

[20] Jiajin Li, Jianheng Tang, Lemin Kong, Huikang Liu, Jia Li, Anthony Man-Cho So, and Jose Blanchet. A convergent single-loop algorithm for relaxation of gromov-wasserstein in graph data. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*, 2023.

[21] Wenqiang Pu, Shahana Ibrahim, Xiao Fu, and Mingyi Hong. Stochastic mirror descent for low-rank tensor decomposition under non-euclidean losses. *IEEE Transactions on Signal Processing*, 70:1803–1818, 2022.

[22] R. Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

[23] Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable Gromov-Wasserstein learning for graph partitioning and matching. In *Advances in Neural Information Processing Systems 32*, 2019.

[24] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. Gromov-Wasserstein learning for graph matching and node embedding. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, pages 6932–6941. PMLR, 2019.

[25] Long Yang, Yu Zhang, Gang Zheng, Qian Zheng, Pengfei Li, Jianhang Huang, and Gang Pan. Policy optimization with stochastic mirror descent. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI 2022)*, volume 36, pages 8823–8831, 2022.

[26] Hui Zhang, Yu-Hong Dai, Lei Guo, and Wei Peng. Proximal-like incremental aggregated gradient method with linear convergence under Bregman distance growth conditions. *Mathematics of Operations Research*, 46(1):61–81, 2021.

[27] Junyu Zhang. Stochastic bergman proximal gradient method revisited: Kernel conditioning and painless variance reduction. *arXiv preprint arXiv:2401.03155*, 2024.

[28] Siqi Zhang and Niao He. On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1806.04781*, 2018.

[29] Daoli Zhu, Sien Deng, Minghua Li, and Lei Zhao. Level-set subdifferential error bounds and linear convergence of Bregman proximal gradient method. *Journal of Optimization Theory and Applications*, 189(3):889–918, 2021.

The appendix is structured as follows. Firstly, in Appendix A, we present some basic definitions essential for our subsequent discussions. Following that, in Appendix B, we offer an exhaustive verification of our assumptions and prove the well-posedness of Algorithm 1 under them. In Appendix C, we introduce the extended Bregman stationarity measure as a tool to investigate (Q) and unveil spurious stationary points. In Appendix D, we provide the proof of the continuity of extended stationarity measure. Then, we furnish the proofs of main results in Appendix E. Finally, we give several examples of spurious stationary points in Appendix F.

Appendix A. Supplementary definitions

We begin by revisiting two types of subgradients for convex functions, crucial for analyzing the optimality condition.

Definition 10 (Subgradients of convex functions) *Rockafellar and Wets [22, Definition 8.3]) For a convex function $l : \mathbb{R}^n \to \overline{\mathbb{R}}$, we define the subgradient at the point $\overline{\boldsymbol{x}}$ by*

$$\partial l(\overline{\boldsymbol{x}}) := \left\{ \boldsymbol{v} \in \mathbb{R}^n : l(\boldsymbol{x}) - l(\overline{\boldsymbol{x}}) \geq \boldsymbol{v}^\top (\boldsymbol{x} - \bar{\boldsymbol{x}}), \text{ for all } \boldsymbol{x} \in \operatorname{dom}(l) \right\},$$

and the horizon subgradient at the point $\overline{\boldsymbol{x}}$ by

$$\partial^\infty l(\overline{\boldsymbol{x}}) := \left\{ \boldsymbol{v} \in \mathbb{R}^n : \lambda^k \boldsymbol{v}^k \to \boldsymbol{v} \text{ with } \lambda^k \to 0^+, \boldsymbol{v}^k \in \partial l(\boldsymbol{x}^k), \text{ and } \boldsymbol{x}^k \to \overline{\boldsymbol{x}} \right\}.$$

Then, we introduce the normal cone to handle the convex constraint.

Definition 11 (Normal cone of convex sets) *[22, Theorem 6.9]) For a convex set $\mathcal{D} \subseteq \mathbb{R}^n$, we define the normal cone at the point $\overline{\boldsymbol{x}} \in \mathcal{D}$ via*

$$\mathcal{N}_{\mathcal{D}}(\overline{\boldsymbol{x}}) := \left\{ \boldsymbol{v} \in \mathbb{R}^n : \boldsymbol{v}^\top (\boldsymbol{x} - \overline{\boldsymbol{x}}) \leq 0, \text{ for all } \boldsymbol{x} \in \mathcal{D} \right\}.$$

For more properties of ∂l, $\partial^\infty l$, and $\mathcal{N}_{\mathcal{D}}$, we refer interested readers to [22, Chapter 8].

Let us revisit the three-point identity for Bregman divergence, as outlined in Lemma 3.1 of [11], i.e.,

$$D_\varphi(z, x) + D_\varphi(x, y) - D_\varphi(z, y) = (z - x) \cdot (\varphi'(y) - \varphi'(x)), \text{ for all } x, y, z \in \operatorname{dom}(\varphi).$$

Since the strict convexity of φ implies the strict monotonic increase of φ', the following fact holds.

Fact 1 *For a strictly convex function φ, the following statements hold:*

(i) *If $z \leq x \leq y$, then $D_\varphi(z, x) \leq D_\varphi(z, y)$.*

(ii) *If $z \leq x \leq y$, then $D_\varphi(x, y) \leq D_\varphi(z, y)$.*

Next, we introduce the notion of the supercoercive property, commonly employed to ensure the well-posedness of the BPG algorithm; see [3, 9].

Definition 12 (Supercoercive property) *Given a function $q : \mathbb{R}^n \to \overline{\mathbb{R}}$, we say that q is supercoercive if for all sequences $\{x^k\}_{k \in \mathbb{N}}$ such that $\|\boldsymbol{x}^k\| \to \infty$, it holds that*

$$\lim_{k \to \infty} \frac{q(\boldsymbol{x}^k)}{\|\boldsymbol{x}^k\|} \to +\infty.$$

Finally, we give the definition of Bregman proximal mapping, which serves as a crucial tool in connecting our stationarity measure with existing ones.

Definition 13 (Bregman proximal mapping) [4, 19] *For $t > 0$ and a kernel function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$, the Bregman proximal mapping for $F : \operatorname{int}(\operatorname{dom}(h)) \to \overline{\mathbb{R}}$ is defined by*

$$\operatorname{prox}_{h,F}^{t}(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{y}\in\mathbb{R}^n} \left\{ F(\boldsymbol{y}) + \frac{1}{t} D_h(\boldsymbol{y}, \boldsymbol{x}) \right\}.$$

Appendix B. Justification of Assumptions

This section is devoted to justifying our assumptions. We first present commonly used kernels that meet Definition 1 to demonstrate the generality of the separable kernels.

Example 1

(i) *Boltzmann–Shannon entropy kernel* $h(\boldsymbol{x}) = \sum_{i=1}^{n} x_i \log(x_i)$;

(ii) *Fermi–Dirac entropy kernel* $h(\boldsymbol{x}) = \sum_{i=1}^{n} x_i \log(x_i) + (1 - x_i)\log(1 - x_i)$;

(iii) *Burg entropy kernel* $h(\boldsymbol{x}) = \sum_{i=1}^{n} -\log(x_i)$;

(iv) *Fractional power kernel* $h(\boldsymbol{x}) = \sum_{i=1}^{n} px_i - \frac{x_i^p}{1-p}$ $(0 < p < 1)$;

(v) *Hellinger entropy kernel* $h(\boldsymbol{x}) = \sum_{i=1}^{n} -\sqrt{1 - x_i^2}$.

The implication of the separable structure is that $\operatorname{cl}(\operatorname{dom}(h))$ forms a box, i.e.,

$$\operatorname{cl}(\operatorname{dom}(h)) = [a, c] \times [a, c] \times \cdots \times [a, c],$$

where $a, c \in \mathbb{R} \cup \{\pm\infty\}$ with $\operatorname{cl}(\operatorname{dom}(\varphi)) = [a, c]$. Due to the convexity of φ and Definition 1 (ii), φ' is monotonically increasing and further $\varphi'(x) \to -\infty$ (resp. $+\infty$) if $x \to a^+$ (resp. $x \to c^-$) and $a > -\infty$ (resp. $c < +\infty$).

B.1. Verification of Assumption 2 (iv)

In this subsection, we aim to identify the condition for Assumption 2 to hold.

(i) When $\operatorname{dom}(\phi)$ is open, e.g., $\phi(x) = 1/x$, to ensure the well-posedness of the Bregman proximal-type algorithms, we should invoke the compactness of $\mathcal{X}$ as stated in Assumption 2 since the condition (2.1) typically fails. Such a supplement is also made in classical Bregman literature; see condition (i) in [3, Lemma 2].

(ii) Even without $\mathcal{X}$ being compact, the closeness of $\operatorname{dom}(\phi)$ and the supercoercivity of local update model is able to ensure that Assumption 2 (iv) is met, under Assumptions 1 and 2 (i)–(iii), see Proposition 14 for further details.

(iii) Without Assumption 2 (iv), the update (1.1) may not have an optimal solution, as illustrated by Example 2.

Example 2 (Necessity of Assumption 2 (iv)) *Let $h(x) = 1/x$, $g(x) = \delta_{\mathbb{R}_+}(x)$, $f(x) = (x-3)^2$, and $\gamma(y;x) = (x-3)^2 + 2(x-3)(y-x)$. Assumption 1 and Assumptions 2 (i)–(iii) hold for this problem with any $\bar{t} > 0$. We can verify that Assumption 2 (iv) fails when we pick $x^k = \frac{1}{2} + \frac{1}{2^k}$, $y^k = 2^k$ and $t = 1$.* It turns out that the well-posedness of (1.1) does not hold. For all $x \in [\frac{1}{2}, 1]$ and $t = 1$, we have

$$\begin{aligned}&\operatorname*{argmin}_{y\in\mathbb{R}_+} \left\{\gamma(y;x) + \frac{1}{t}D_h(y,x)\right\}\\ =&\operatorname*{argmin}_{y\in\mathbb{R}_+} \left\{\frac{1}{y} + \left(2x + \frac{1}{x^2} - 6\right)(y-x)\right\} = \varnothing,\end{aligned}$$

where the last equality is due to $2x + \frac{1}{x^2} - 6 \leq -1$ for $x \in [\frac{1}{2}, 1]$.

Proposition 14 *Suppose that $\mathrm{dom}(\phi)$ is closed. The following hold:*

(i) *If the surrogate model takes the form $\gamma(\boldsymbol{y};\boldsymbol{x}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x})$ and $h + tg$ is supercoercive for all $t > 0$, then Assumption 2 (iv) is satisfied.*

(ii) *If the surrogate model takes the form $\gamma(\boldsymbol{y};\boldsymbol{x}) = f(\boldsymbol{y})$ and $h + tF$ is supercoercive for all $t > 0$, then Assumption 2 (iv) is satisfied.*

(iii) *If the surrogate model takes the form $\gamma(\boldsymbol{y};\boldsymbol{x}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{x})^\top\nabla^2 f(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})$, $h + tg$ is supercoercive for all $t > 0$, and f is a convex function, then Assumption 2 (iv) holds.*

Remark 15 *Regarding (iii), the convexity condition for f is also posited in Doikov and Nesterov [13].*

Proof Case (i):

To justify that Assumption 2 (iv) is satisfied, we proceed to prove a stronger counterpart, i.e.,

$$\lim_{k\to\infty} \frac{\gamma(\boldsymbol{y}^k;\boldsymbol{x}^k) + g(\boldsymbol{y}^k) + \frac{1}{t}D_h(\boldsymbol{y}^k,\boldsymbol{x}^k)}{\|\boldsymbol{y}_k\|} = +\infty.$$

When $\|\boldsymbol{y}^k\| \to \infty$, $\boldsymbol{x}^k \to \overline{\boldsymbol{x}}$, and continuity of ∇f, we have

$$\lim_{k\to\infty} \frac{\gamma(\boldsymbol{y}^k;\boldsymbol{x}^k)}{\|\boldsymbol{y}^k\|} = \nabla f(\overline{\boldsymbol{x}})^\top \lim_{k\to\infty} \frac{\boldsymbol{y}^k}{\|\boldsymbol{y}^k\|} < +\infty.$$

Subsequently, it remains to show

$$\lim_{k\to\infty} \frac{\frac{1}{t}D_h(\boldsymbol{y}^k,\boldsymbol{x}^k) + g(\boldsymbol{y}^k)}{\|\boldsymbol{y}^k\|} = +\infty.$$

To do so, we consider the interior coordinates and boundary coordinates separately. Recall that $\mathcal{B}(\bar{\boldsymbol{x}}) = \{b \in [n] : \overline{x}_b \in \mathrm{bd}(\mathrm{dom}(\varphi))\}$ and $\mathcal{I}(\bar{\boldsymbol{x}}) = \{i \in [n] : \overline{x}_i \in \mathrm{int}(\mathrm{dom}(\varphi))\}$. Without loss of generality (WLOG), we assume that $\mathrm{dom}(\varphi) = [a, b]$.

(a) For all $i \in \mathcal{I}(\bar{\boldsymbol{x}})$, we have $\lim_{k\to\infty} |\varphi'(x_i^k)| = |\varphi'(\bar{x}_i)| < \infty$ due to the continuous differentiability of φ. Then, we have

$$\begin{aligned}\lim_{k\to\infty} \frac{D_\varphi(y_i^k, x_i^k)}{\|\boldsymbol{y}^k\|} &= \lim_{k\to\infty} \frac{\varphi(y_i^k) - \varphi(x_i^k) - \varphi'(x_i^k)(y_i^k - x_i^k)}{\|\boldsymbol{y}^k\|}, \\ &= \lim_{k\to\infty} \frac{\varphi(y_i^k)}{\|\boldsymbol{y}^k\|} - \varphi'(\bar{x}_i) \cdot \lim_{k\to\infty} \frac{y_i^k}{\|\boldsymbol{y}^k\|}, \quad \forall i \in \mathcal{I}(\bar{\boldsymbol{x}}), \end{aligned} \tag{B.1}$$

where the first equality follows from the definition of D_φ and the second equality is due to the finiteness of $\varphi(\bar{x}_i)$ and $\bar{x}_i$.

(b) For all $b \in \mathcal{B}(\bar{\boldsymbol{x}})$, we may assume WLOG that $\bar{x}_b = a$. As $\lim_{k\to\infty} x_b^k = \bar{x}_b = a$, we have $x_b^k < x_b^0$ for sufficiently large k. Then, we try to give a lower bound on $D_\varphi(y_b^k, x_b^k)$ and analyze its limit for all $b \in \mathcal{B}(\bar{\boldsymbol{x}})$.

If $y_b^k \le x_b^0$, then we have $D_\varphi(y_b^k, x_b^0) \le D_\varphi(a, x_b^0)$ due to Fact 1 (ii) and $a < y_b^k \le x_b^0$. It follows that

$$D_\varphi(y_b^k, x_b^k) \ge 0 \ge D_\varphi(y_b^k, x_b^0) - D_\varphi(a, x_b^0), \quad \forall b \in \mathcal{B}(\bar{\boldsymbol{x}}).$$

Otherwise, we have

$$D_\varphi(y_b^k, x_b^k) \ge D_\varphi(y_b^k, x_b^0), \quad \forall b \in \mathcal{B}(\bar{\boldsymbol{x}}),$$

due to Fact 1 (i) and $x_b^k < x_b^0 < y_b^k$. Consequently, we get

$$\lim_{k\to\infty} \frac{D_\varphi(y_b^k, x_b^k)}{\|\boldsymbol{y}^k\|} \ge \lim_{k\to\infty} \frac{D_\varphi(y_b^k, x_b^0)}{\|\boldsymbol{y}^k\|} = \lim_{k\to\infty} \frac{\varphi(y_b^k) - \varphi'(x_b^0) y_b^k}{\|\boldsymbol{y}^k\|}, \quad \forall b \in \mathcal{B}(\bar{\boldsymbol{x}}), \tag{B.2}$$

where the first inequality follows from the finiteness of $D_\varphi(a, x_b^0)$.

Together (B.1) and (B.2), we have

$$\begin{aligned}&\lim_{k\to\infty} \frac{\frac{1}{t} D_h(\boldsymbol{y}^k, \boldsymbol{x}^k) + g(\boldsymbol{y}^k)}{\|\boldsymbol{y}^k\|} \\ \ge &\lim_{k\to\infty} \frac{\frac{1}{t} h(\boldsymbol{y}^k) + g(\boldsymbol{y}^k)}{\|\boldsymbol{y}^k\|} - \frac{1}{t} \sum_{i\in\mathcal{I}(\bar{\boldsymbol{x}})} \varphi'(\bar{x}_i) \cdot \lim_{k\to\infty} \frac{y_i^k}{\|\boldsymbol{y}^k\|} - \frac{1}{t} \sum_{b\in\mathcal{B}(\bar{\boldsymbol{x}})} \varphi'(x_b^0) \cdot \lim_{k\to\infty} \frac{y_b^k}{\|\boldsymbol{y}^k\|} = +\infty,\end{aligned}$$

where the equality holds since $h + tf$ is supercoercive for all $t > 0$ and the finiteness of $\varphi'(x_b^0)$ and $\varphi'(\bar{x}_i)$ for all $b \in \mathcal{B}(\bar{\boldsymbol{x}})$ and $i \in \mathcal{I}(\bar{\boldsymbol{x}})$. This completes the proof of case (i).

Case (ii): The ideas of the proof of case (ii) are similar to those of case (i). We just need to replace g with F.

Case (iii): Given the convexity of f, we find that $(\boldsymbol{y} - \boldsymbol{x})^\top \nabla^2 f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) \ge 0$ for any $\boldsymbol{y} \in \mathbb{R}^n$. Thus, it suffices to show

$$\lim_{k\to\infty} f(\boldsymbol{x}^k) + \nabla f(\boldsymbol{x}^k)^T (\boldsymbol{y}^k - \boldsymbol{x}^k) + g(\boldsymbol{y}^k) + \frac{1}{t} D_h(\boldsymbol{y}^k, \boldsymbol{x}^k) = +\infty,$$

which reduces to the case (i). We complete the proof. ■

B.2. Proof of Lemma 2

With Assumptions 1 and 2 justified, we prove Lemma 2 to establish the well-posedness of Algorithm 1 under them.

Proof We prove the result by contradiction. Suppose that $\boldsymbol{x} \in \operatorname{int}(\operatorname{dom}(h))$ and $T_\gamma^t(\boldsymbol{x}) \notin \operatorname{int}(\operatorname{dom}(h))$. Then, can select an interior point $\mathbf{x}^{\text{int}} \in \operatorname{int}(\operatorname{dom}(h)) \cap \mathcal{X}$, where its distance from $T_\gamma^t(\mathbf{x})$ is bounded. Subsequently, we construct an intermediate point as

$$\boldsymbol{x}^\theta := \theta \boldsymbol{x}^{\text{int}} + (1-\theta) T_\gamma^t(\boldsymbol{x}),$$

and an univariate function as

$$\phi(\theta) := \gamma\left(\boldsymbol{x}^\theta; \boldsymbol{x}\right) + \frac{1}{t} D_h\left(\boldsymbol{x}^\theta, \boldsymbol{x}\right) + g\left(\boldsymbol{x}^\theta\right),$$

where $\theta \in (0, 1]$.

Our strategy is to show that

$$\lim_{\theta \to 0_+} \frac{\phi(\theta) - \phi(0)}{\theta} = -\infty$$

holds [1]. Then, it is sufficient to demonstrate that there exists a constant $\theta \in (0, 1)$ such that $\phi(\theta) - \phi(0)$, thereby contradicting the definition of $T_\gamma^t(\boldsymbol{x})$.

Now, we focus on the decomposition of $\phi(\theta) - \phi(0)$ as below:

$$\begin{aligned}
&\phi(\theta) - \phi(0) \\
&= \gamma\left(\boldsymbol{x}^\theta; \boldsymbol{x}\right) - \gamma(T_\gamma^t(\boldsymbol{x}); \boldsymbol{x}) + g\left(\boldsymbol{x}^\theta\right) - g(T_\gamma^t(\boldsymbol{x})) + \frac{1}{t}\left(D_h\left(\boldsymbol{x}^\theta, \boldsymbol{x}\right) - D_h(T_\gamma^t(\boldsymbol{x}), \boldsymbol{x})\right) \\
&= \left[\gamma\left(\boldsymbol{x}^\theta; \boldsymbol{x}\right) - \gamma(T_\gamma^t(\boldsymbol{x}); \boldsymbol{x})\right] + \left[g\left(\boldsymbol{x}^\theta\right) - g(T_\gamma^t(\boldsymbol{x}))\right] + \left[\frac{1}{t}\sum_{j=1}^n \left(\varphi\left(x_j^\theta\right) - \varphi\left(T_\gamma^t(\boldsymbol{x})_j\right)\right)\right] \\
&\quad \left[-\frac{\theta}{t}\sum_{j=1}^n \varphi'(x_j)\left(x_j^{\text{int}} - T_\gamma^t(\boldsymbol{x})_j\right)\right].
\end{aligned}$$

Then, we address each component individually.

(i) Due to the differentiability of $\gamma(\cdot, \boldsymbol{x})$ at the point $\boldsymbol{x}$ (see Assumption 2 (ii)), there exists a sufficient small θ such that

$$\left|\frac{\gamma\left(\boldsymbol{x}^\theta; \boldsymbol{x}\right) - \gamma\left(T_\gamma^t(\boldsymbol{x}); \boldsymbol{x}\right)}{\theta}\right| \leq \mathcal{O}(1).$$

(ii) Given the local Lipschitz continuity of g on $\mathcal{X}$, there also exists a sufficiently small θ (i.e., you can certainly choose the same θ to satisfy both (i) and (ii)) such that

$$\left|\frac{g\left(\boldsymbol{x}^\theta\right) - g(T_\gamma^t(\boldsymbol{x}))}{\theta}\right| \leq \mathcal{O}(1).$$

1. $\theta \to 0^+$ denotes the limit of θ as it approaches 0 from the right.

(iii) By Mean Value Theorem, we have

$$\sum_{j=1}^{n} \frac{\varphi\left(x_j^\theta\right) - \varphi\left(T_\gamma^t(\boldsymbol{x})_j\right)}{\theta} = \sum_{j=1}^{n} \varphi'\left(z_j^\theta\right) \cdot \left(x_j^{\text{int}} - T_\gamma^t(\boldsymbol{x})_j\right),$$

where z_j^θ is in the interval between x_j^θ and $T_\gamma^t(\boldsymbol{x})_j$ for all $j \in [n]$. When $\theta \to 0_+$, the term $(x_j^{\text{int}} - T_\gamma^t(\boldsymbol{x})_j)$ is always bounded. Moreover, we have $z_b^\theta \to T_\gamma^t(\boldsymbol{x})_b$ when $\theta \to 0^+$. Thus, the key lies in the boundedness of $\varphi'(T_\gamma^t(\boldsymbol{x})_j))$ for all $j \in [n]$.

We will discuss the interior coordinates and boundary coordinate of $T_\gamma^t(\boldsymbol{x})$ separately.

(a) For all $i \in \mathcal{I}(T_\gamma^t(\boldsymbol{x}))$, we know that $T_\gamma^t(\mathbf{x})_i \in \text{int}(\text{dom}(\varphi))$, and therefore $\varphi'(T_\gamma^t(\mathbf{x})_i)$ is bounded due to the continuity of φ'.

(b) For all $b \in \mathcal{B}(T_\gamma^t(\boldsymbol{x}))$, we have $T_\gamma^t(\boldsymbol{x})_b \in \text{bd}(\text{dom}(\varphi))$ and thus $|\varphi'(T_\gamma^t(\boldsymbol{x})_b)| = +\infty$ from Definition 1 (ii). WLOG, we can assume that $\text{cl}(\text{dom}(\varphi)) = [a, c]$ and $T_\gamma^t(\boldsymbol{x})_b = a$. Moreover, as φ' is strictly increasing by the strict convexity of φ, we have $\varphi'(T_\gamma^t(\boldsymbol{x})_b) = -\infty$. Together with the fact $x_b^{\text{int}} - T_\gamma^t(\boldsymbol{x})_b > 0$, we have

$$\lim_{\theta \to 0_+} \varphi'\left(z_b^\theta\right) \cdot \left(x_b^{\text{int}} - T_\gamma^t(\boldsymbol{x})_b\right) = -\infty.$$

(iv) The final term is constant w.r.t. θ and thus bounded for sure.

Putting all the pieces together, we conclude the proof. ■

Appendix C. Extended Bregman stationarity measure

To investigate (Q) and develop tools for studying spurious stationary points, we define an extended Bregman stationarity measure in this section.

To begin, we unify existing stationarity measures as $R_\gamma^t(\boldsymbol{x}) := D_h(T_\gamma^t(\boldsymbol{x}), \boldsymbol{x})$,[2] where $t > 0$ is the step size. Conceptually, we use the relative change under the Bregman geometry to quantify the stationarity, which has been well explored in the literature [7, 15, 16, 18]. If we set $\gamma = f$, then the update mapping $T_\gamma^t(\boldsymbol{x})$ coincides with $\text{prox}_{h,F}^t(\boldsymbol{x})$, and thus R_γ^t recovers the stationarity gap $D_h(\text{prox}_{h,F}^t(\boldsymbol{x}), \boldsymbol{x})$ proposed by [28].

Despite the unification, R_γ^t is still not well-defined on the boundary $\text{bd}(\text{dom}(h))$, as the mapping $\boldsymbol{x} \mapsto T_\gamma^t(\boldsymbol{x})$ defined by (1.1) involves the Bregman divergence function $(\boldsymbol{y}, \boldsymbol{x}) \mapsto D_h(\boldsymbol{y}, \boldsymbol{x})$, which is only defined on $\text{dom}(h) \times \text{int}(\text{dom}(h))$. Given this limitation, we are motivated to extend the domain of Bregman stationarity measures to $\text{cl}(\text{dom}(h))$. As a preliminary action, we define the extended update mapping that is applicable to $\text{cl}(\text{dom}(h)) \cap \mathcal{X}$ as follows.

Definition 16 *We define the extended update mapping by* $\overline{T}_\gamma^t(\boldsymbol{x}) := \underset{\boldsymbol{y} \in \mathcal{X}}{\text{argmin}}\ G_\gamma^t(\boldsymbol{y}; \boldsymbol{x})$, *where*

$$G_\gamma^t(\boldsymbol{y}; \boldsymbol{x}) := \gamma(\boldsymbol{y}; \boldsymbol{x}) + g(\boldsymbol{y}) + \underbrace{\frac{1}{t} \sum_{i \in \mathcal{I}(\boldsymbol{x})} D_\varphi(y_i, x_i)}_{\textit{Interior coordinates}} + \underbrace{\delta_{\boldsymbol{y}_{\mathcal{B}(\boldsymbol{x})} = \boldsymbol{x}_{\mathcal{B}(\boldsymbol{x})}}(\boldsymbol{y})}_{\textit{Boundary coordinates}}.$$

2. Here, we ignore the difference between $D_h(T_\gamma^t(\boldsymbol{x}), \boldsymbol{x})$ and $D_h(\boldsymbol{x}, T_\gamma^t(\boldsymbol{x}))$ for simplicity. As we will discuss in Remark 29, this difference would not affect our theoretical results.

This newly updated rule distinguishes between interior and boundary coordinates. For boundary coordinates of $\boldsymbol{x}$, we enforce $\boldsymbol{y}$ to be equal to $\boldsymbol{x}$, while for interior coordinates, we update following the original update rule (1.1). This construction enables us to focus on the boundary points to address non-gradient Lipschitz kernel functions.

Armed with the extended update mapping $\overline{T}^t_\gamma$, we are ready to provide the formal definition of the extended Bregman stationarity measure $\overline{R}^t_\gamma$. It is worth noting that $\overline{R}^t_\gamma$ is defined over the entire domain $\mathcal{X}$, and it retrieves $R^t_\gamma(\mathbf{x})$ for $\boldsymbol{x} \in \operatorname{int}(\operatorname{dom}(h)) \cap \mathcal{X}$.

Definition 17 (Extended stationarity measure) *We define the extended Bregman stationarity measure* $\overline{R}^t_\gamma(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}$ *as* $\overline{R}^t_\gamma(\boldsymbol{x}) := \sum_{i \in \mathcal{I}(\boldsymbol{x})} D_\varphi(\overline{T}^t_\gamma(\boldsymbol{x})_i, x_i)$.

C.1. Well-definedness

To prove the well-definedness of the extended measure $\overline{R}^t_\gamma$, it suffices to prove the well-definedness of the extended mapping $\overline{T}^t_\gamma$. For this purpose, we first prove that the level sets of $G^t_\gamma(\,\cdot\,;\boldsymbol{x})$ are nonempty; see Lemma 18. Then, due to Assumptions 2 (i) and (iv), we see that the level sets of $G^t_\gamma(\,\cdot\,;\boldsymbol{x})$ are compact. Further, Assumption 2 (iii) guarantees the strict convexity of $G^t_\gamma(\,\cdot\,;\boldsymbol{x})$. Both the strict convexity and level boundedness of G^t_γ ensure the well-definedness and uniqueness of the extended update mapping $\overline{T}^t_\gamma$.

Lemma 18 (Existence) *For all* $\boldsymbol{x} \in \mathcal{X}$*, we have*

$$\operatorname{argmin}_{\boldsymbol{y} \in \mathcal{X}} \; G^t_\gamma(\boldsymbol{y};\boldsymbol{x}) \neq \varnothing.$$

Proof To ensure the existence of the optimal solution, it suffices to show that, for some constant $c \in \mathbb{R}$, the level set $\{\boldsymbol{y} \in \mathcal{X} : G^t_\gamma(\boldsymbol{y};\boldsymbol{x}) \leq c\}$ is compact for any $\boldsymbol{x} \in \mathcal{X}$. At first, from [22, Theorem 1.6], we know this level set is closed due to the lower semicontinuity of : $G^t_\gamma(\cdot,\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathcal{X}$.

Next, our task is to show the set $\{\boldsymbol{y} \in \mathcal{X} : G^t_\gamma(\boldsymbol{y};\boldsymbol{x}) \leq c\}$ is bounded. We consider two cases separately:

(i) If $\boldsymbol{x} \in \operatorname{int}(\operatorname{dom}(h)) \cap \mathcal{X}$, the boundedness of the set $\{\boldsymbol{y} \in \mathcal{X} : G^t_\gamma(\boldsymbol{y};\boldsymbol{x}) \leq c\}$ is from Assumption 2 (iv), as the coerciveness can imply one non-empty bounded level set.

(ii) If $\boldsymbol{x} \in \operatorname{bd}(\operatorname{dom}(h)) \cap \mathcal{X}$, it suffices to show the coerciveness of $G^t_\gamma(\cdot;\boldsymbol{x})$. That is, if we consider an arbitrary sequence $\{\boldsymbol{y}^k\} \subseteq \mathcal{X}$ with $\|\boldsymbol{y}^k\| \to +\infty$, we have $G^t_\gamma(\boldsymbol{y}^k;\boldsymbol{x}) \to +\infty$.

WLOG, we can assume that $\boldsymbol{y}^k_{\mathcal{B}(\boldsymbol{x})} \equiv \boldsymbol{x}_{\mathcal{B}(\boldsymbol{x})}$; otherwise, $G^t_\gamma(\boldsymbol{y}^k;\boldsymbol{x}) = +\infty$, for any $k \in \mathbb{N}$. Then, we pick up an interior point $\boldsymbol{x}^{\text{int}} \in \operatorname{int}(\operatorname{dom}(h)) \cap \mathcal{X}$ and construct three sequences $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$, $\{\tilde{\boldsymbol{y}}^k\}_{k\in\mathbb{N}}$ and $\{\tilde{\boldsymbol{x}}^k\}_{k\in\mathbb{N}}$ satisfying

$$\begin{aligned}
&\boldsymbol{x}^k \to \boldsymbol{x} \text{ with } \mathcal{B}(\boldsymbol{x}^k) = \mathcal{B}(\boldsymbol{x}), \forall k \in \mathbb{N}\\
&\tilde{\boldsymbol{y}}^k = (1-\theta_k)\boldsymbol{y}^k + \theta_k \boldsymbol{x}^{\text{int}}, \forall k \in \mathbb{N},\\
&\tilde{\boldsymbol{x}}^k = (1-\theta_k)\boldsymbol{x}^k + \theta_k \boldsymbol{x}^{\text{int}}, \forall k \in \mathbb{N},
\end{aligned}$$

where $\theta_k \in (0,1)$ and $\theta_k \to 0$. Then, it is trivial to conclude that $\|\tilde{\boldsymbol{y}}^k\| \to +\infty$ and $\tilde{\boldsymbol{x}}^k \to \boldsymbol{x}$. Moreover, due to $\tilde{\boldsymbol{y}}^k_{\mathcal{B}(\boldsymbol{x})} = \tilde{\boldsymbol{x}}^k_{\mathcal{B}(\boldsymbol{x})}, \forall k \in \mathbb{N}$, we get

$$\gamma(\tilde{\boldsymbol{y}}^k;\tilde{\boldsymbol{x}}^k) + g(\tilde{\boldsymbol{y}}^k) + \frac{1}{t}D_h(\tilde{\boldsymbol{y}}^k,\tilde{\boldsymbol{x}}^k) = \gamma(\tilde{\boldsymbol{y}}^k;\tilde{\boldsymbol{x}}^k) + g(\tilde{\boldsymbol{y}}^k) + \frac{1}{t}\sum_{i\in\mathcal{I}(\boldsymbol{x})} D_\varphi(\tilde{y}^k_i,\tilde{x}^k_i).$$

By the continuity of the mappings $\gamma(\cdot), g(\cdot)$, and $D_\phi(\cdot,\cdot)$, there exists a sufficiently small $\theta_k > 0$ such that

$$\left|\gamma(\tilde{\boldsymbol{y}}^k;\tilde{\boldsymbol{x}}^k) + g(\tilde{\boldsymbol{y}}^k) + \frac{1}{t}D_h(\tilde{\boldsymbol{y}}^k,\tilde{\boldsymbol{x}}^k) - G^t_\gamma(\boldsymbol{y}^k;\boldsymbol{x})\right| \le 1.$$

From Assumption 2 (iv), we know $\gamma(\tilde{\boldsymbol{y}}^k;\tilde{\boldsymbol{x}}^k) + g(\tilde{\boldsymbol{y}}^k) + \frac{1}{t}D_h(\tilde{\boldsymbol{y}}^k,\tilde{\boldsymbol{x}}^k) \to +\infty$. Thus, we can conclude that $G^t_\gamma(\boldsymbol{y}^k;\boldsymbol{x}) \to +\infty$.

■

C.2. Continuity of extended stationarity measure

To investigate the equivalence described in (Q), our first step is to establish the continuity of $\overline{R}^t_\gamma$. This task involves recognizing that $\overline{R}^t_\gamma$'s definition depends on both $\overline{T}^t_\gamma$ and $\mathcal{I}(\boldsymbol{x})$, which may exhibit discontinuity. To start, we establish the continuity of the extended update mapping:

Proposition 19 *The extended update mapping $\overline{T}^t_\gamma : \mathcal{X} \to \mathbb{R}^n$ is continuous on the domain $\mathcal{X}$.*

The continuity of $\overline{T}^t_\gamma$ on $\mathcal{X}$ serves as a fundamental property of the extended update mapping. Not only does it provide insight into $\overline{T}^t_\gamma(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathrm{bd}(\mathrm{dom}(h)) \cap \mathcal{X}$, but it also plays a crucial role in establishing the continuity of $\overline{R}^t_\gamma$. Leveraging the continuity of $\overline{T}^t_\gamma$ and the structure of $\mathcal{I}(\boldsymbol{x})$, we establish one of the main theoretical results as below:

Theorem 20 *The extended stationarity measure $\overline{R}^t_\gamma : \mathcal{X} \to \mathbb{R}$ is continuous on the domain $\mathcal{X}$.*

C.3. Necessity of zero extended stationarity measure

In this subsection, we will establish the necessary conditions by utilizing the extended stationarity measure. The key idea for proving necessity is through the fixed-point equation $\overline{T}^t_\gamma(\boldsymbol{x}) = \boldsymbol{x}$. Below is the flowchart outlining the proofs:

$$\boxed{\mathbf{0} \in \partial F(\boldsymbol{x})} \overset{\text{Proposition 22}}{\Longrightarrow} \boxed{\overline{T}^t_\gamma(\boldsymbol{x}) = \boldsymbol{x}} \overset{\text{Proposition 21}}{\Longleftrightarrow} \boxed{\overline{R}^t_\gamma(\boldsymbol{x}) = 0} \text{ for } \boldsymbol{x} \in \mathcal{X}.$$

We start with establishing the equivalence between $\overline{T}^t_\gamma(\boldsymbol{x}) = \boldsymbol{x}$ and $\overline{R}^t_\gamma(\boldsymbol{x}) = 0$.

Proposition 21 *For all $\boldsymbol{x} \in \mathcal{X}$, the extended stationarity measure being zero, i.e., $\overline{R}^t_\gamma(\boldsymbol{x}) = 0$, is equivalent to $\overline{T}^t_\gamma(\boldsymbol{x}) = \boldsymbol{x}$.*

Proof Considering the definition of $\overline{T}^t_\gamma(\boldsymbol{x})$, we observe that $\overline{T}^t_\gamma(\boldsymbol{x})_{\mathcal{B}(\boldsymbol{x})} = \boldsymbol{x}_{\mathcal{B}(\boldsymbol{x})}$ on the boundary coordinates. Regarding those interior coordinates, the definition of $\overline{R}^t_\gamma$ guarantees that $\overline{R}^t_\gamma(\boldsymbol{x}) = 0$ if and only if $\overline{T}^t_\gamma(\boldsymbol{x})_{\mathcal{I}(\boldsymbol{x})} = \boldsymbol{x}_{\mathcal{I}(\boldsymbol{x})}$. By combining these two observations and the fact that $\mathcal{I}(\boldsymbol{x}) \cup \mathcal{B}(\boldsymbol{x}) = [n]$, we establish the equivalence between $\overline{R}^t_\gamma(\boldsymbol{x}) = 0$ and $\overline{T}^t_\gamma(\boldsymbol{x}) = \boldsymbol{x}$. ■

Next, we will investigate the relation between the fixed-point equation $\overline{T}^t_\gamma(\boldsymbol{x}) = \boldsymbol{x}$ and the stationary condition $\mathbf{0} \in \partial F(\boldsymbol{x})$.

Proposition 22 *If $\boldsymbol{x} \in \mathcal{X}$ is a stationary point, then we have $\overline{T}^t_\gamma(\boldsymbol{x}) = \boldsymbol{x}$.*

Proof To start, it is sufficient to demonstrate that $\boldsymbol{x} = \operatorname{argmin}_{\boldsymbol{y}} G^t_\gamma(\boldsymbol{y}; \boldsymbol{x})$, a condition that is equivalent to $\mathbf{0} \in \partial G^t_\gamma(\boldsymbol{x}; \boldsymbol{x})$ according to optimality condition. Based on Assumption 2 (ii) and [22, Corollary 10.9], we have

$$\partial G^t_\gamma(\boldsymbol{y}; \boldsymbol{x}) \mid_{\boldsymbol{y}=\boldsymbol{x}} = \nabla f(\boldsymbol{x}) + \partial g(\boldsymbol{x}) + \partial \delta_{\boldsymbol{y}_{\mathcal{B}(\boldsymbol{x})} = \boldsymbol{x}_{\mathcal{B}(\boldsymbol{x})}}(\boldsymbol{y}) \mid_{\boldsymbol{y}=\boldsymbol{x}} . \tag{C.1}$$

Since we have $\mathbf{0} \in \partial \delta_{\boldsymbol{y}_{\mathcal{B}(\boldsymbol{x})} = \boldsymbol{x}_{\mathcal{B}(\boldsymbol{x})}}(\boldsymbol{y}) \mid_{\boldsymbol{y}=\boldsymbol{x}}$, we know $\mathbf{0} \in \partial F(\boldsymbol{x})$ and then $\mathbf{0} \in \partial G^t_\gamma(\boldsymbol{x}; \boldsymbol{x})$. ■

Equipped with Proposition 21 and Proposition 22, we are poised to demonstrate our primary discovery: The extended stationarity measure equals zero for all stationary points.

Theorem 23 *If $\boldsymbol{x} \in \mathcal{X}$ is a stationary point, i.e., $\mathbf{0} \in \partial F(\boldsymbol{x})$, then we have $\overline{R}^t_\gamma(\boldsymbol{x}) = 0$.*

Combining Theorem 23 with the continuity of $\overline{R}^t_\gamma$ (as demonstrated in Theorem 20), we can now establish the necessity direction of equivalence (Q).

Corollary 24 *Let the sequence $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}} \subseteq \operatorname{int}(\operatorname{dom}(h)) \cap \mathcal{X}$ converge to $\overline{\boldsymbol{x}} \in \mathcal{X}$ with $\mathbf{0} \in \partial F(\overline{\boldsymbol{x}})$. Then, we have*

$$\lim_{k\to\infty} \overline{R}^t_\gamma(\boldsymbol{x}^k) = 0. \tag{C.2}$$

Proof Here is a one-line proof:

$$\lim_{k\to\infty} \overline{R}^t_\gamma(\boldsymbol{x}^k) = \overline{R}^t_\gamma\left(\lim_{k\to\infty} \boldsymbol{x}^k\right) = \overline{R}^t_\gamma(\overline{\boldsymbol{x}}) = 0,$$

where the first equality follows from the continuity of $\overline{R}^t_\gamma$ (see Theorem 20), and the last one is due to the neccessity of $\overline{R}^t_\gamma$ (see Theorem 23). ■

Corollary 24 claims that if a limit point of a sequence is stationary, then (C.2) holds true. However, existing counterparts of (C.2) are typically established under the assumption that the kernel functions are gradient Lipschitz, see [21, 25, 28]. In contrast to existing works, Corollary 24 presented in this paper can be applied to a broader class of kernel functions, such as those without the gradient Lipschitz property.

C.4. Non-sufficiency of zero extended stationarity measure

Based on the extended stationarity measure, we give a characterization of spurious stationary points, demostrating the non-sufficiency of zero extended stationarity measure. Then, the proof of Proposition 5 directly follows.

Proposition 25 (Characterization of spurious stationary points) *A point $\boldsymbol{x} \in \mathcal{X}$ is a spurious stationary point if and only if*

$$\overline{R}_\gamma^t(\boldsymbol{x}) = 0 \text{ but } \mathbf{0} \notin \partial F(\boldsymbol{x}).$$

Proof From Definition 3, it is sufficient to show the equivalence between $\overline{R}_\gamma^t(\boldsymbol{x}) = 0$ and the existence of a vector $\boldsymbol{p} \in \partial F(\boldsymbol{x})$ where $\boldsymbol{p}_{\mathcal{I}(\boldsymbol{x})} = 0$. Following the proof of Proposition 22, we will proceed to check the equivalent optimality condition of $\mathbf{0} \in \partial G_\gamma^t(\boldsymbol{x}; \boldsymbol{x})$. Observing that $\partial \delta_{\boldsymbol{y}_{\mathcal{B}(\boldsymbol{x})} = \boldsymbol{x}_{\mathcal{B}(\boldsymbol{x})}}(\boldsymbol{x}) = \operatorname{span}\{\mathbf{e}_b : b \in \mathcal{B}(\boldsymbol{x})\}$, we know that $\mathbf{0} \in \partial G_\gamma^t(\boldsymbol{x}; \boldsymbol{x})$ is equivalent to the existence of a vector $\boldsymbol{p} \in \partial F(\boldsymbol{x})$ with $\boldsymbol{p}_{\mathcal{I}(\boldsymbol{x})} = \mathbf{0}$. We complete the proof. ■

Considering the continuity of $\overline{R}_\gamma^t$, if a sequence $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}} \subseteq \mathcal{X} \cap \operatorname{int}(\operatorname{dom}(h))$ converges to a spurious stationary point, then $\overline{R}_\gamma^t(\boldsymbol{x}^k) \to 0$. This proves proves Proposition 5.

Appendix D. Proof of continuity of extended stationarity measure

To begin, we introduce two technical lemmas essential for proving Proposition 19.

Lemma 26 *Suppose the sequence $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}} \subseteq \operatorname{int}(\operatorname{dom}(h)) \cap \mathcal{X}$ is bounded, and the sequence $\{T_\gamma^t(\boldsymbol{x}^k)\}_{k\in\mathbb{N}}$ converges to $\boldsymbol{x}^* \in \mathcal{X}$. We define a sequence $\{\boldsymbol{d}^{k+1}\}_{k\in\mathbb{N}}$ satisfying*

$$\nabla\gamma\left(T_\gamma^t(\boldsymbol{x}^k); \boldsymbol{x}^k\right) + \boldsymbol{d}^{k+1} + \frac{1}{t}\left(\nabla h\left(T_\gamma^t(\boldsymbol{x}^k)\right) - \nabla h(\boldsymbol{x}^k)\right) = \mathbf{0} \text{ and } \boldsymbol{d}^{k+1} \in \partial g\left(T_\gamma^t(\boldsymbol{x}^k)\right). \quad \text{(D.1)}$$

Then, the sequence $\{\boldsymbol{d}^{k+1}\}_{k\in\mathbb{N}}$ is bounded.

Proof We prove the boundedness of $\{\boldsymbol{d}^{k+1}\}_{k\in\mathbb{N}}$ by contradiction. Suppose that the sequence $\{\boldsymbol{d}^{k+1}\}_{k\in\mathbb{N}}$ is not bounded. Then, there must exist a subsequence that diverges. WLOG, we can assume that $\|\boldsymbol{d}^{k+1}\| \to \infty$ and $\boldsymbol{d}^{k+1}/\|\boldsymbol{d}^{k+1}\| \to \boldsymbol{d}^*$ for some $\boldsymbol{d}^*$. Due to Definition 10 and $T_\gamma^t(\boldsymbol{x}^k) \to \boldsymbol{x}^*$, we have $\boldsymbol{d}^* \in \partial^\infty g(\boldsymbol{x}^*)$. Moreover, owing to the convexity and continuity of the function g, we can apply Rockafellar and Wets [22, Proposition 8.12] to get $\partial^\infty g(\boldsymbol{x}^*) = \mathcal{N}_{\mathcal{X}}(\boldsymbol{x}^*)$ and hence we have

$$(\boldsymbol{x}^* - \boldsymbol{x})^\top \boldsymbol{d}^* \geq 0 \text{ for all } \boldsymbol{x} \in \mathcal{X}. \quad \text{(D.2)}$$

In view of (D.2), we consider the following two scenarios separately:

(i) $(\boldsymbol{x}^* - \boldsymbol{x})^\top \boldsymbol{d}^* = 0$ for all $\boldsymbol{x} \in \mathcal{X}$;

(ii) There exists some $\boldsymbol{x} \in \mathcal{X}$ such that $(\boldsymbol{x}^* - \boldsymbol{x})^\top \boldsymbol{d}^* > 0$.

Scenarios (i): For all $k \in \mathbb{N}$, we have

$$(T_\gamma^t(\boldsymbol{x}^k) - \boldsymbol{x}^k)^\top \boldsymbol{d}^* = 0, \quad \text{(D.3)}$$

by substituting $\boldsymbol{x} = \boldsymbol{x}^k$ and $\boldsymbol{x} = T_\gamma^t(\boldsymbol{x}^k)$ in (D.2), and summing them.

From (D.3), we have

$$\varphi'\left(T_\gamma^t(\boldsymbol{x}^k)_i\right) - \varphi'(x_i^k) = -t\left(\nabla_i\gamma(T_\gamma^t(\boldsymbol{x}^k);\boldsymbol{x}^k) + d_i^{k+1}\right), \forall i \in [n]. \tag{D.4}$$

Next, we want to argue that the left-hand side of (D.4) will go to infinity, which will contradict with (D.3).

Given the boundedness of the sequence $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$ and $\{T_\gamma^t(\boldsymbol{x}^k)\}_{k\in\mathbb{N}} \to \boldsymbol{x}^*$, it follows from the joint continuity property of $\nabla\gamma(\cdot;\cdot)$, as stated in Assumption 2 (i), that $\{\nabla_i\gamma(T_\gamma^t(\mathbf{x}^k);\mathbf{x}^k)\}_{k\in\mathbb{N}}$ is also bounded. We proceed to discuss the term d_i^{k+1} for all $i \in [n]$. Let $\mathcal{I}^+ := \{i \in [n] : d_i^* > 0\}$ and $\mathcal{I}^- := \{i \in [n] : d_i^* < 0\}$. Thus, we have $d_i^{k+1} \to +\infty$ (resp. $d_i^{k+1} \to -\infty$) by $d_i^{k+1}/\|\boldsymbol{d}^{k+1}\| \to d_i^*$, for all $i \in \mathcal{I}^+$ (resp. $\mathcal{I}^-$). Hence, the equation (D.4) yields

$$\varphi'\left(T_\gamma^t(\boldsymbol{x}^k)_i\right) - \varphi'(x_i^k) \to -\infty \text{ (resp. } +\infty) \text{ for } i \in \mathcal{I}^+ \text{ (resp. } i \in \mathcal{I}^-).$$

Moreover, since φ' is strictly increasing by the strict convexity of φ, for sufficiently large k, we have

$$T_\gamma^t(\boldsymbol{x}^k)_i < x_i^k \text{ for } i \in \mathcal{I}^+ \text{ and } T_\gamma^t(\boldsymbol{x}^k)_i > x_i^k \text{ for } i \in \mathcal{I}^-.$$

Consequently, we get

$$(T_\gamma^t(\boldsymbol{x}^k) - \boldsymbol{x}^k)^\top \boldsymbol{d}^* < 0 \text{ for sufficiently large } k,$$

which contradicts (D.3).

Scenarios (ii): There must exists a point $x \in \mathcal{X}$ and a positive constant $\alpha > 0$ such that $(\boldsymbol{x}^* - \boldsymbol{x})^\top \boldsymbol{d}^* \geq \alpha$. Additionally, we can select some $\boldsymbol{x}^{\text{int}} \in \text{int}(\text{dom}(h)) \cap \mathcal{X}$, which is sufficiently close to $\boldsymbol{x}$ such that $(\boldsymbol{x}^{\text{int}} - x)^T \boldsymbol{d}^* \leq \frac{\alpha}{2}$. Therefore, we have

$$(\boldsymbol{x}^* - \boldsymbol{x}^{\text{int}})^T \boldsymbol{d}^* = \lim_{k\to\infty} (T_\gamma^t(\boldsymbol{x}^k) - \boldsymbol{x}^{\text{int}})^T \frac{\boldsymbol{d}^{k+1}}{\|\boldsymbol{d}^{k+1}\|} \geq \frac{\alpha}{2}.$$

where the first equality follows from $T_\gamma^t(\boldsymbol{x}^k) \to \boldsymbol{x}^*$ and the definition of $\boldsymbol{d}^*$. Then, for sufficiently large k, we have

$$(T_\gamma^t(\boldsymbol{x}^k) - \boldsymbol{x}^{\text{int}})^T \boldsymbol{d}^{k+1} \geq \frac{\alpha}{2}\|\boldsymbol{d}^{k+1}\|. \tag{D.5}$$

Furthermore, by multiplying (D.1) by $T_\gamma^t(\boldsymbol{x}^k) - \boldsymbol{x}^{\text{int}}$, we obtain

$$(T_\gamma^t(\boldsymbol{x}^k) - \boldsymbol{x}^{\text{int}})^\top \left[\nabla\gamma\left(T_\gamma^t(\boldsymbol{x}^k);\boldsymbol{x}^k\right) + \boldsymbol{d}^{k+1} + \frac{1}{t}\left(\nabla h(T_\gamma^t(\boldsymbol{x}^k)) - \nabla h(\boldsymbol{x}^k)\right)\right] = 0.$$

It follows that

$$\begin{aligned}
&(T_\gamma^t(\boldsymbol{x}^k) - \boldsymbol{x}^{\text{int}})^\top \left(\nabla h(T_\gamma^t(\boldsymbol{x}^k)) - \nabla h(\boldsymbol{x}^k)\right) \\
=& -(T_\gamma^t(\boldsymbol{x}^k) - \boldsymbol{x}^{\text{int}})^T \boldsymbol{d}^{k+1} - (T_\gamma^t(\boldsymbol{x}^k) - \boldsymbol{x}^{\text{int}})^T \nabla\gamma\left(T_\gamma^t(\boldsymbol{x}^k);\boldsymbol{x}^k\right) \\
\leq& -\frac{\alpha}{2}\|\boldsymbol{d}^{k+1}\|^2 - (T_\gamma^t(\boldsymbol{x}^k) - \boldsymbol{x}^{\text{int}})^T \nabla\gamma\left(T_\gamma^t(\boldsymbol{x}^k);\boldsymbol{x}^k\right) \to -\infty,
\end{aligned} \tag{D.6}$$

where the first inequality is due to (D.5), and the last limit follows from the boundedness of $\{\nabla\gamma(T_\gamma^t(\boldsymbol{x}^k);\boldsymbol{x}^k)\}_{k\in\mathbb{N}}$ and $\|\boldsymbol{d}^{k+1}\| \to \infty$.

We now utilize (D.6) to establish the contradiction. The key proof idea follows from the one in Lemma 2. For all $k \in \mathbb{N}$, we construct a sequence of intermediate points as

$$\boldsymbol{x}^{\theta,k} := \theta \boldsymbol{x}^{\text{int}} + (1-\theta) T_\gamma^t(\boldsymbol{x}^k),$$

and a sequence of univariate functions as

$$\phi_k(\theta) := \gamma\left(\boldsymbol{x}^{\theta,k}; \boldsymbol{x}\right) + \frac{1}{t} D_h\left(\boldsymbol{x}^{\theta,k}, \boldsymbol{x}\right) + g\left(\boldsymbol{x}^{\theta,k}\right),$$

where $\theta \in (0,1]$. By the definition of $T_\gamma^t(\boldsymbol{x}^k)$, we have $\phi_k(0) = \min_{\theta\in[0,1]} \phi_k(\theta)$ for $k \in \mathbb{N}$. Then, our strategy is to show that for sufficiently large k

$$\lim_{\theta \to 0_+} \frac{\phi_k(\theta) - \phi_k(0)}{\theta} < 0,$$

and thus yields a contradiction. By Mean Value Theorem, we have

$$\begin{aligned}
&\frac{\phi_k(\theta) - \phi_k(0)}{\theta} \\
=&\frac{\gamma\left(\boldsymbol{x}^{\theta,k}; \boldsymbol{x}\right) - \gamma\left(T_\gamma^t(\boldsymbol{x}^{\theta,k}); \boldsymbol{x}\right)}{\theta} + \frac{1}{t}(T_\gamma^t(\boldsymbol{x}^k) - \boldsymbol{x}^{\text{int}})^\top \left(\nabla h(\boldsymbol{z}^{\theta,k}) - \nabla h(\boldsymbol{x}^k)\right) \\
&+ \frac{g\left(\boldsymbol{x}^{\theta,k}\right) - g\left(T_\gamma^t(\boldsymbol{x}^k)\right)}{\theta},
\end{aligned} \tag{D.7}$$

where $\boldsymbol{z}^{\theta,k}$ is between $\boldsymbol{x}^{\theta,k}$ and $T_\gamma^t(\boldsymbol{x}^k)$. Owing to the continuous differentiability of γ and local Lipschitz continuity of g, we know the first and third terms in (D.7) are bounded. From (D.6), as θ approaches 0_+, the second term tends toward $-\infty$ for sufficiently large k. Consequently, we have

$$\lim_{k\to\infty} \lim_{\theta \to 0^+} \frac{\phi_k(\theta) - \phi_k(0)}{\theta} = -\infty,$$

which leads to a contradiction. We complete our proof. ■

Next, we extend Lemma 26 to a more general case, whose proof techiniques are essentially the same as the one developed in Lemma 26.

Corollary 27 *Suppose that the sequence $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}} \subseteq \mathcal{X}$ is bounded satisfying $\mathcal{I}(\boldsymbol{x}^k) \equiv \mathcal{I}_0 \subseteq [n]$ and $\mathcal{B}(\boldsymbol{x}^k) \equiv \mathcal{B}_0$, and the sequence $\{\overline{T}_\gamma^t(\boldsymbol{x}^k)\}_{k\in\mathbb{N}}$ converges to $\boldsymbol{x}^* \in \mathcal{X}$. We define a sequence $\{\boldsymbol{d}^{k+1}\}_{k\in\mathbb{N}}$ satisfying*

$$\nabla\gamma(\overline{T}_\gamma^t(\boldsymbol{x}^k); \boldsymbol{x}^k) + \boldsymbol{d}^{k+1} + \boldsymbol{p}^{k+1} + \frac{1}{t}\sum_{i\in\mathcal{I}_0} \nabla\left(D_\varphi(\overline{T}_\gamma^t(\boldsymbol{x}^k)_i, x_i^k)\right) = \boldsymbol{0}, \tag{D.8}$$

where $\boldsymbol{d}^{k+1} \in \partial g\left(\overline{T}_\gamma^t(\boldsymbol{x}^k)\right)$ and $\boldsymbol{p}^{k+1} \in \partial\delta_{\boldsymbol{y}_{\mathcal{B}_0}=\boldsymbol{x}_{\mathcal{B}_0}^k}\left(\overline{T}_\gamma^t(\boldsymbol{x}^k)\right)$, for all $k \in \mathbb{N}$. Then, the sequence $\{\boldsymbol{d}_{\mathcal{I}_0}^{k+1}\}_{k\in\mathbb{N}}$ is bounded.

Proof WLOG, we assume that $\mathrm{cl}(\mathrm{dom}(\varphi)) = [a, c]$ and $\boldsymbol{x}^k_{\mathcal{B}_0} \equiv \boldsymbol{x}^0_{\mathcal{B}_0}$ for all $k \in \mathbb{N}$. Let $\tilde{g} = g + \delta_{\boldsymbol{y}_{\mathcal{B}_0} = \boldsymbol{x}^0_{\mathcal{B}_0}}$, $\tilde{\mathcal{X}} = \mathcal{X} \cap \{\boldsymbol{y} \in \mathcal{X} : \boldsymbol{y}_{\mathcal{B}_0} = \boldsymbol{x}^0_{\mathcal{B}_0}\}$, and $\tilde{\boldsymbol{d}}^{k+1} = \boldsymbol{d}^{k+1} + \boldsymbol{p}^{k+1}$. Then, we have

$$\tilde{\boldsymbol{d}}^{k+1}_{\mathcal{B}_0} = \boldsymbol{d}^{k+1}_{\mathcal{B}_0},\ \tilde{\boldsymbol{d}}^{k+1} \in \partial \tilde{g}\left(\overline{T}^t_\gamma(\boldsymbol{x}^k)\right),\ \text{and}\ \mathrm{dom}(\tilde{g}) = \tilde{\mathcal{X}}.$$

We continue to rewrite (D.8) as

$$\nabla \gamma(\overline{T}^t_\gamma(\boldsymbol{x}^k); \boldsymbol{x}^k) + \tilde{\boldsymbol{d}}^{k+1} + \frac{1}{t}\sum_{i \in \mathcal{I}_0} \nabla\left(D_\varphi(\overline{T}^t_\gamma(\boldsymbol{x}^k)_i, x^k_i)\right) = \boldsymbol{0}, \tag{D.9}$$

where $\tilde{\boldsymbol{d}}^{k+1} \in \partial \tilde{g}\left(\overline{T}^t_\gamma(\boldsymbol{x}^k)\right)$. Due to the boundedness of $\{\overline{T}^t_\gamma(\boldsymbol{x}^k)\}_{k\in\mathbb{N}}$ and $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$, and the continuity of $\nabla\gamma$, it follows that the sequence $\{\nabla\gamma(\overline{T}^t_\gamma(\boldsymbol{x}^k); \boldsymbol{x}^k)\}_{k\in\mathbb{N}}$ is bounded. Moreover, we know the sequence $\{\tilde{\boldsymbol{d}}^{k+1}_{\mathcal{B}_0}\}_{k\in\mathbb{N}}$ is bounded from the boundedness of $\{\overline{T}^t_\gamma(\boldsymbol{x}^k)\}_{k\in\mathbb{N}}$ and (D.9).

We left to show the sequence $\{\boldsymbol{d}^{k+1}_{\mathcal{I}_0}\}_{k\in\mathbb{N}}$ is bounded. We prove this result by contradiction. Suppose that the sequence $\{\boldsymbol{d}^{k+1}_{\mathcal{I}_0}\}_{k\in\mathbb{N}}$ is unbounded. Hence, the sequence $\{\tilde{\boldsymbol{d}}^{k+1}\}_{k\in\mathbb{N}}$ is also unbounded. Then, there must exist a subsequence that diverges. WLOG, we can assume that $\tilde{\boldsymbol{d}}^{k+1}/\|\tilde{\boldsymbol{d}}^{k+1}\| \to \overline{\boldsymbol{d}}$ for some $\overline{\boldsymbol{d}}$. Here, we have $\|\overline{\boldsymbol{d}}_{\mathcal{I}_0}\| = 1$ and $\overline{\boldsymbol{d}}_{\mathcal{B}_0} = \boldsymbol{0}$ due to boundedness of $\{\tilde{\boldsymbol{d}}^{k+1}_{\mathcal{B}_0}\}_{k\in\mathbb{N}}$. Then, the left proof is the same as the one in Lemma 26. We omit the proof details. ■

Lemma 28 *If the sequence $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}} \subseteq \mathrm{int}(\mathrm{dom}(h)) \cap \mathcal{X}$ is bounded, then the sequence $\{T^t_\gamma(\boldsymbol{x}^k)\}_{k\in\mathbb{N}}$ is also bounded.*

Proof We prove this result by contradiction. WLOG, we can assume $\boldsymbol{x}^k \to \overline{\boldsymbol{x}} \in \mathcal{X}$ by the boundedness of $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$. From Assumption 2 (iv), we have

$$\lim_{k\to+\infty} \gamma(T^t_\gamma(\boldsymbol{x}^k), \boldsymbol{x}^k) + \frac{1}{t} D_h(T^t_\gamma(\boldsymbol{x}^k), \boldsymbol{x}^k) + g(T^t_\gamma(\boldsymbol{x}^k)) = +\infty. \tag{D.10}$$

Moreover, we have

$$\lim_{k\to+\infty} \gamma(\boldsymbol{x}^k; \boldsymbol{x}^k) + \frac{1}{t} D_h(\boldsymbol{x}^k, \boldsymbol{x}^k) + g(\boldsymbol{x}^k) = \gamma(\overline{\boldsymbol{x}}; \overline{\boldsymbol{x}}) + g(\overline{\boldsymbol{x}}) = f(\overline{\boldsymbol{x}}) + g(\overline{\boldsymbol{x}}) < +\infty,$$

where the first equality is owing to the continuity of γ and g, the second one is from Assumption 2 (ii), and the last inequality is due to $\overline{\boldsymbol{x}} \in \mathcal{X}$. Then, together with (D.10), for k large enough, it follows that

$$\gamma(T^t_\gamma(\boldsymbol{x}^k), \boldsymbol{x}^k) + \frac{1}{t} D_h(T^t_\gamma(\boldsymbol{x}^k), \boldsymbol{x}^k) + g(T^t_\gamma(\boldsymbol{x}^k)) > \gamma(\boldsymbol{x}^k, \boldsymbol{x}^k) + \frac{1}{t} D_h(\boldsymbol{x}^k, \boldsymbol{x}^k) + g(\boldsymbol{x}^k),$$

which contradicts with the definition of $T^t_\gamma(\boldsymbol{x}^k)$. Hence, the sequence $\{T^t_\gamma(\boldsymbol{x}^k)\}_{k\in\mathbb{N}}$ is bounded. ■

Now, we are ready to give a full proof of Proposition 19.

D.1. Proof of Proposition 19

Proof To show that the mapping $\overline{T}^t_\gamma(\cdot)$ is continuous, it suffices to show that for any sequence $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}} \subseteq \mathcal{X}$ converging to $\boldsymbol{x} \in \mathcal{X}$, it holds that:

$$\lim_{k\to\infty} \overline{T}^t_\gamma(\boldsymbol{x}^k) = \overline{T}^t_\gamma(\boldsymbol{x}).$$

To proceed, we consider the following two scenarios sequentially:

(i) The sequence $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}} \subseteq \mathrm{int}(\mathrm{dom}(h)) \cap \mathcal{X}$.

(ii) The sequence $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}} \subseteq \mathrm{cl}(\mathrm{dom}(h)) \cap \mathcal{X}$.

To start with, we consider the simple case (i). Later on, we can extend our proof strategy to consider the general case by considering the interior and boundary coordinates of $\boldsymbol{x}$ separately. Then, we can reduce the general case to the simple case considered here.

Scenarios (i): Due to $\overline{T}^t_\gamma(\cdot) = T^t_\gamma(\cdot)$ on $\mathrm{int}(\mathrm{dom}(h)) \cap \mathcal{X}$. It is equivalent to show

$$\overline{T}^t_\gamma(\boldsymbol{x}) = \lim_{k\to\infty} T^t_\gamma(\boldsymbol{x}^k).$$

The remainder of the proof proceeds in two steps. We give a proof sketch here initially.

- **Step 1**: For the boundary coordinates of $\boldsymbol{x}$, we have $\lim_{k\to\infty} T^t_\gamma(\boldsymbol{x}^k)_{\mathcal{B}(\boldsymbol{x})} = \overline{T}^t_\gamma(\boldsymbol{x})_{\mathcal{B}(\boldsymbol{x})}$.

By Lemma 28, we can pass to a subsequence such that $T^t_\gamma(\boldsymbol{x}^k) \to \tilde{\boldsymbol{x}} \in \mathrm{cl}(\mathrm{dom}(h))$. Then, we have to show $\tilde{\boldsymbol{x}}_{\mathcal{B}(\boldsymbol{x})} = \boldsymbol{x}_{\mathcal{B}(\boldsymbol{x})}$ from the definition of $\overline{T}^t_\gamma$, i.e., $\overline{T}^t_\gamma(\boldsymbol{x})_{\mathcal{B}(\boldsymbol{x})} = \boldsymbol{x}_{\mathcal{B}(\boldsymbol{x})}$. We prove this result by contradiction. Our proof strategy essentially follows the one we developed in Lemma 2 and Lemma 26. For all $k \in \mathbb{N}$, we construct a sequence of intermediate points as

$$\boldsymbol{x}^{\theta,k} := \theta\boldsymbol{x}^k + (1-\theta)T^t_\gamma(\boldsymbol{x}^k),$$

and a sequence of univariate functions as

$$\phi_k(\theta) := \gamma\left(\boldsymbol{x}^{\theta,k}; \boldsymbol{x}\right) + \frac{1}{t} D_h\left(\boldsymbol{x}^{\theta,k}, \boldsymbol{x}\right) + g\left(\boldsymbol{x}^{\theta,k}\right),$$

where $\theta \in (0,1]$. Then, we show $\phi_k(\theta) < \phi_k(0)$ would hold for some $\theta \in (0,1]$ and $k \in \mathbb{N}$ if $\tilde{\boldsymbol{x}}_{\mathcal{B}(\boldsymbol{x})} \neq \boldsymbol{x}_{\mathcal{B}(\boldsymbol{x})}$.

- **Step 2**: We prove $\tilde{\boldsymbol{x}} = \lim_{k\to\infty} T^t_\gamma(\boldsymbol{x}^k) = \overline{T}^t_\gamma(\boldsymbol{x})$ via the optimality condition.

For all $k \in \mathbb{N}$, from the definition of $T^t_\gamma(\boldsymbol{x}^k)$, we can write down its optimality condition:

$$\nabla\gamma\left(T^t_\gamma(\boldsymbol{x}^k); \boldsymbol{x}^k\right) + \frac{1}{t}\left(\nabla h(T^t_\gamma(\boldsymbol{x}^k)) - \nabla h(\boldsymbol{x}^k)\right) + \boldsymbol{d}^{k+1} = \boldsymbol{0}, \tag{D.11}$$

where $\boldsymbol{d}^{k+1} \in \partial g(T^t_\gamma(\boldsymbol{x}^k))$. As $\boldsymbol{x}_k \to \boldsymbol{x}$ and $T^t_\gamma(\boldsymbol{x}^k) \to \tilde{\boldsymbol{x}}$, we can apply Lemma 26 and get the boundedness of $\{\boldsymbol{d}^{k+1}\}_{k\in\mathbb{N}}$. By passing to a subsequence, we can assume that $\boldsymbol{d}^{k+1} \to \overline{\boldsymbol{d}}$. Then,

we have $\overline{\boldsymbol{d}} \in \partial g(\tilde{\boldsymbol{x}})$. As k approaches infinity in (D.11), and given that the limit of $\nabla h(\boldsymbol{x}^k)$ exists for the coordinates corresponding to $\mathcal{I}(\boldsymbol{x})$, it follows that

$$\nabla_{\mathcal{I}(\boldsymbol{x})}\gamma\left(\tilde{\boldsymbol{x}};\boldsymbol{x}\right)+\frac{1}{t}\left(\nabla_{\mathcal{I}(\boldsymbol{x})}h(\tilde{\boldsymbol{x}})-\nabla_{\mathcal{I}(\boldsymbol{x})}h(\boldsymbol{x})\right)+\overline{\boldsymbol{d}}_{\mathcal{I}(\boldsymbol{x})}=\boldsymbol{0}. \tag{D.12}$$

Moreover, from **Step 1**, we know $\tilde{\boldsymbol{x}}_{\mathcal{B}(\boldsymbol{x})}=\boldsymbol{x}_{\mathcal{B}(\boldsymbol{x})}$ and thus (D.12) is the optimality condition of the problem

$$\min_{\boldsymbol{y}\in\mathbb{R}^n} G_\gamma^t(\boldsymbol{y};\boldsymbol{x})=\gamma(\boldsymbol{y};\boldsymbol{x})+\delta_{\boldsymbol{y}_{\mathcal{B}(\boldsymbol{x})}=\boldsymbol{x}_{\mathcal{B}(\boldsymbol{x})}}(\boldsymbol{y})+\frac{1}{t}\sum_{i\in\mathcal{I}(\boldsymbol{x})}D_\varphi(y_i,x_i)+g(\boldsymbol{y}). \tag{D.13}$$

From Definition 16, we know $\tilde{\boldsymbol{x}}=\overline{T}_\gamma^t(\boldsymbol{x})$.

Scenarios (ii): The key steps essentially follow those developed in **Scenarios (i)**. We highlight the key differences and omit redundant details for simplicity.

To prove **Step 1**, we have to show the sequence $\{\overline{T}_\gamma^t(\boldsymbol{x}^k)\}_{k\in\mathbb{N}}$ is bounded, as we we did in Lemma 28. We prove this result by contradiction. By passing to a subsequence, we assume that $\|\overline{T}_\gamma^t(\boldsymbol{x}^k)\|\to+\infty$. Then, we construct a sequence $\{\boldsymbol{y}^k\}_{k\in\mathbb{N}}\subseteq\operatorname{int}(\operatorname{dom}(h))\cap\mathcal{X}$ that satisfying $\|\boldsymbol{x}^k-\boldsymbol{y}^k\|\le 2^{-k}$ and $\|T_\gamma^t(\boldsymbol{y}^k)-\overline{T}_\gamma^t(\boldsymbol{x}^k)\|\le 1$. Such a sequence exists due to the result of **Scenarios (i)** and the existence of interior points. That is, for each $k\in\mathbb{N}$, we can always construct a sequence converging to $\boldsymbol{x}^k$, and pick a point $\boldsymbol{y}^k$ in the sequence satisfying the conditions. Consequently, we have $\|T_\gamma^t(\boldsymbol{y}^k)\|\to+\infty$ when $\boldsymbol{y}^k\to\boldsymbol{x}$. We obtain the contradiction by applying the same arguments in Lemma 28.

From the boundedness of the sequence $\{\overline{T}_\gamma^t(\boldsymbol{x}^k)\}_{k\in\mathbb{N}}$, we can assume $\overline{T}_\gamma^t(\boldsymbol{x}^k)\to\tilde{\boldsymbol{x}}\in\operatorname{cl}(\operatorname{dom}(h))$ by passing to a subsequence if necessary. For all $k\in\mathbb{N}$, we construct a sequence of intermediate points as

$$\overline{\boldsymbol{x}}^{\theta,k}:=\theta\boldsymbol{x}^k+(1-\theta)\overline{T}_\gamma^t(\boldsymbol{x}^k),$$

and a sequence of univariate functions as

$$\overline{\phi}_k(\theta):=G_\gamma^t\left(\overline{\boldsymbol{x}}^{\theta,k};\boldsymbol{x}^k\right),$$

where $\theta\in(0,1]$. We can prove that $\overline{T}_\gamma^t(\boldsymbol{x})_{\mathcal{B}(\boldsymbol{x})}=\tilde{\boldsymbol{x}}_{\mathcal{B}(\boldsymbol{x})}=\boldsymbol{x}_{\mathcal{B}(\boldsymbol{x})}$ by the same arguments developed in **Step 1** for **Scenarios (i)**.

Step 2: Owing to $\boldsymbol{x}_k\to\boldsymbol{x}$, WLOG, we can assume that $\mathcal{I}(\boldsymbol{x}^k)=\mathcal{I}_0\subseteq[n]$. From the definition of $\overline{T}_\gamma^t(\boldsymbol{x}^k)$, we have

$$\nabla_{\mathcal{I}_0}\gamma\left(\overline{T}_\gamma^t(\boldsymbol{x}^k);\boldsymbol{x}^k\right)+\frac{1}{t}\left(\nabla_{\mathcal{I}_0}h(\overline{T}_\gamma^t(\boldsymbol{x}^k))-\nabla_{\mathcal{I}_0}h(\boldsymbol{x}^k)\right)+\boldsymbol{d}_{\mathcal{I}_0}^{k+1}=\boldsymbol{0}, \tag{D.14}$$

where $\boldsymbol{d}^{k+1}\in\partial g(\overline{T}_\gamma^t(\boldsymbol{x}^k))$. Moreover, the sequence $\{\boldsymbol{d}_{\mathcal{I}_0}^{k+1}\}_{k\in\mathbb{N}}$ is bounded due to Corollary 27. Then, there always exists a subsequence such that $\boldsymbol{d}_{\mathcal{I}_0}^{k+1}\to\overline{\boldsymbol{d}}_{\mathcal{I}_0}$ for some $\overline{\boldsymbol{d}}\in\mathbb{R}^n$. As k approaches infinity in (D.14), and given that the limit of $\nabla h(\boldsymbol{x}^k)$ exists for the coordinates corresponding to $\mathcal{I}(\boldsymbol{x})$, it follows that

$$\nabla_{\mathcal{I}(\boldsymbol{x})}\gamma\left(\tilde{\boldsymbol{x}};\boldsymbol{x}\right)+\frac{1}{t}\left(\nabla_{\mathcal{I}(\boldsymbol{x})}h(\tilde{\boldsymbol{x}})-\nabla_{\mathcal{I}(\boldsymbol{x})}h(\boldsymbol{x})\right)+\overline{\boldsymbol{d}}_{\mathcal{I}(\boldsymbol{x})}=\boldsymbol{0}.$$

Lastly, due to $\tilde{\boldsymbol{x}}_{\mathcal{B}(\boldsymbol{x})} = \boldsymbol{x}_{\mathcal{B}(\boldsymbol{x})}$, we know $\overline{T}_\gamma^t(\boldsymbol{x}) = \tilde{\boldsymbol{x}}$.

For the final section, our objective is to furnish a comprehensive proof within **Step 1** under **Scenarios (i)**. To demonstrate that $\phi_k(\theta) < \phi_k(0)$ holds for some $\theta \in (0,1]$ and $k \in \mathbb{N}$ if $\tilde{\boldsymbol{x}}_{\mathcal{B}(\boldsymbol{x})} \neq \boldsymbol{x}_{\mathcal{B}(\boldsymbol{x})}$, we decompose $\phi_k(\theta) - \phi_k(0)$ following the proof outlined in Lemma 2 and 26. By Mean Value Theorem, we have

$$\begin{aligned}&\frac{\phi_k(\theta) - \phi_k(0)}{\theta}\\ =&\frac{\gamma\left(\boldsymbol{x}^{\theta,k};\boldsymbol{x}\right) - \gamma\left(T_\gamma^t(\boldsymbol{x}^{\theta,k});\boldsymbol{x}\right)}{\theta} + \frac{1}{t}(T_\gamma^t(\boldsymbol{x}^k) - \boldsymbol{x}^k)^\top\left(\nabla h(\boldsymbol{z}^{\theta,k}) - \nabla h(\boldsymbol{x}^k)\right)\\ &+ \frac{g\left(\boldsymbol{x}^{\theta,k}\right) - g\left(T_\gamma^t(\boldsymbol{x}^k)\right)}{\theta},\end{aligned}$$

where $\boldsymbol{z}^{\theta,k}$ is between $\boldsymbol{x}^{\theta,k}$ and $T_\gamma^t(\boldsymbol{x}^k)$. Following the same argument we did in (D.7). The first term and the third term are uniformly bounded. We just focus on the second term here.

Since we have assumed $\tilde{\boldsymbol{x}}_{\mathcal{B}(\boldsymbol{x})} \neq \boldsymbol{x}_{\mathcal{B}(\boldsymbol{x})}$, there is a $b_* \in \mathcal{B}(\boldsymbol{x})$ such that $\tilde{x}_{b_*} \neq x_{b_*}$. As $z_{b_*}^{\theta,k}$ lies in the interval between $x_{b_*}^{\theta,k}$ and $T_\gamma^t(\boldsymbol{x}^k)_{b_*}$, we have $z_{b_*}^{\theta,k} - x_{b_*}^{\theta,k} = \xi \cdot (T_\gamma^t(\boldsymbol{x}^k)_{b_*} - x_{b_*}^k)$ for some $\xi > 0$. The monotone of φ' yields

$$\left(\varphi'\left(z_{b_*}^{\theta,k}\right) - \varphi'(x_{b_*}^k)\right) \cdot \left(x_{b_*}^k - T_\gamma^t(\boldsymbol{x}^k)_{b_*}\right) \leq 0. \tag{D.15}$$

Using $\tilde{x}_{b_*} \neq x_{b_*}$, we have $x_{b_*}^k - T_\gamma^t(\boldsymbol{x}^k)_{b_*} \nrightarrow 0$. Then, noticing $|\varphi'(x_{b_*}^k)| \to \infty$ due to $x_{b_*} \in \mathrm{bd}(\mathrm{dom}(\varphi))$, we have

$$\left(\varphi'\left(z_{b_*}^{\theta,k}\right) - \varphi'(x_{b_*}^k)\right) \cdot \left(x_{b_*}^k - T_\gamma^t(\boldsymbol{x}^k)_{b_*}\right) \to -\infty \text{ as } k \to \infty.$$

Thus, we get

$$\lim_{k\to\infty} \frac{\phi_k(\theta) - \phi_k(0)}{\theta} = -\infty,$$

which yields a contradiction. We complete our proof. ■

D.2. Proof of Theorem 20

Proof For $\boldsymbol{x} \in \mathrm{int}(\mathrm{dom}(h)) \cap \mathcal{X}$, we have $\mathcal{I}(\boldsymbol{x}) = [n]$ and the continuity of $\overline{R}_\gamma^t$ at $\boldsymbol{x}$ follows from the continuity of $\overline{T}_\gamma^t$ at $\boldsymbol{x}$. Then, we only need to show that $\overline{T}_\gamma^t$ is continuous at $\boldsymbol{x} \in \mathrm{bd}(\mathrm{dom}(h))$. That is, for any $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$ converging to $\boldsymbol{x} \in \mathrm{bd}(\mathrm{dom}(h))$,

$$\lim_{k\to\infty} \overline{R}_\gamma^t(\boldsymbol{x}^k) = \overline{R}_\gamma^t(\boldsymbol{x}). \tag{D.16}$$

Recall the definition of $\overline{R}_\gamma^t$ as

$$\overline{R}_\gamma^t(\boldsymbol{x}) = \sum_{i\in\mathcal{I}(\boldsymbol{x})} D_\varphi(\overline{T}_\gamma^t(\boldsymbol{x})_i, x_i).$$

Then, the main difficulty to verifying (D.16) is that $\mathcal{I}(\boldsymbol{x}^k)$ may not equal $\mathcal{I}(\boldsymbol{x})$ for k sufficiently large.

WLOG, we can assume that $\mathcal{I}(\boldsymbol{x}^k) \equiv \mathcal{I}_0 \subseteq [n]$ for all $k \in \mathbb{N}$. Then, we have $\mathcal{I}(\boldsymbol{x}) \subseteq \mathcal{I}_0$ as $\boldsymbol{x}^k \to \boldsymbol{x}$. Now, we discuss $\mathcal{I}(\boldsymbol{x})$ and $\mathcal{I}_0 \setminus \mathcal{I}(\boldsymbol{x})$ separately.

(i) For all $i \in \mathcal{I}(\boldsymbol{x})$, we have

$$D_\varphi\left(\overline{T}_\gamma^t(\boldsymbol{x}^k)_i, x_i^k\right) \to D_\varphi\left(\overline{T}_\gamma^t(\boldsymbol{x})_i, x_i\right),$$

due to the continuity of $\overline{T}_\gamma^t$ and $x_i^k \to x_i \in \operatorname{int}(\operatorname{dom}(\varphi))$.

(ii) For all $i \in \mathcal{I}_0 \setminus \mathcal{I}(\boldsymbol{x})$, we want to show

$$D_\varphi\left(\overline{T}_\gamma^t(\boldsymbol{x}^k)_i, x_i^k\right) \to 0.$$

To see this, we notice that the convexity of φ yields

$$D_\varphi\left(\overline{T}_\gamma^t(\boldsymbol{x}^k)_i, x_i^k\right) \le \left(\varphi'\left(\overline{T}_\gamma^t(\boldsymbol{x}^k)_i\right) - \varphi'(x_i^k)\right) \cdot \left(\overline{T}_\gamma^t(\boldsymbol{x}^k)_i - x_i^k\right).$$

To show the right-hand side goes to zero for $i \in \mathcal{I}_0 \setminus \mathcal{I}(\boldsymbol{x})$, we revisit the optimality condition (D.14). Owing to the boundedness of $\{\boldsymbol{d}_{\mathcal{I}_0}^{k+1}\}_{k\in\mathbb{N}}$ and $\nabla\gamma(\overline{T}_\gamma^t(\boldsymbol{x}^k); \boldsymbol{x}^k)$, it follows that

$$\left|\varphi'\left(\overline{T}_\gamma^t(\boldsymbol{x}^k)_i\right) - \varphi'(x_i^k)\right| \le \mathcal{O}(1).$$

By continuity of $\overline{T}_\gamma^t$ and $\boldsymbol{x}^k \to \boldsymbol{x}$, we have $x_i^k \to x_i$ and $\overline{T}_\gamma^t(\boldsymbol{x}^k)_i \to \overline{T}_\gamma^t(\boldsymbol{x})_i$ for $i \in \mathcal{I}_0 \setminus \mathcal{I}(\boldsymbol{x})$. Note that $\mathcal{I}_0 \setminus \mathcal{I}(\boldsymbol{x}) \subseteq \mathcal{B}(\boldsymbol{x})$ and $\overline{T}_\gamma^t(\boldsymbol{x})_b = x_b$ for $b \in \mathcal{B}(\boldsymbol{x})$ by the definition of $\overline{T}_\gamma^t$. It follows that $\overline{T}_\gamma^t(\boldsymbol{x}^k)_i - x_i^k \to 0$ for $i \in \mathcal{I}_0 \setminus \mathcal{I}(\boldsymbol{x})$, which completes the proof.

Remark 29 *We remark that the non-symmetry of Bregman divergnece, i.e., $D_h(\boldsymbol{y}, \boldsymbol{x}) \neq D_h(\boldsymbol{x}, \boldsymbol{y})$, does not affect our results. For instance, if $R^t(\boldsymbol{x}) = D_h(\boldsymbol{x}, T_\gamma^t(\boldsymbol{x}))$, and correspondingly,*

$$\overline{R}_\gamma^t(\boldsymbol{x}) := \sum_{i\in\mathcal{I}(\boldsymbol{x})} D_\varphi\left(x_i, \overline{T}_\gamma^t(\boldsymbol{x})_i\right),$$

then we still get the continuity of $\overline{R}_\gamma^t$ by applying the proof of Theorem 20 with Bregman divergences $D_\varphi\left(\overline{T}_\gamma^t(\boldsymbol{x}^k)_i, x_i^k\right)$ and $D_\varphi\left(\overline{T}_\gamma^t(\boldsymbol{x})_i, x_i\right)$ replaced by $D_\varphi\left(x_i^k, \overline{T}_\gamma^t(\boldsymbol{x}^k)_i\right)$ and $D_\varphi\left(x_i, \overline{T}_\gamma^t(\boldsymbol{x})_i\right)$.

■

Appendix E. Proof of main results

In this section, we provide the missing proofs of main results in Sec. 3, which are mainly based on the continuity property of the extended stationarity measure and the definition of spurious stationary points. We note that the proof of Proposition 5 has been given in Sec. C.4.

E.1. Proof of Theorem 7

Proof For any $K \in \mathbb{N}$ and arbitrary $\epsilon > 0$, our goal is to construct the initial point $\boldsymbol{x}^0$ sufficiently close to the spurious point $\tilde{\boldsymbol{x}}$ such that $\|\boldsymbol{x}^K - \tilde{\boldsymbol{x}}^*\| \leq \epsilon$.

The key step lies in proving the following claim is correct: Given an arbitrary $\epsilon_0 > 0$, there exist $\epsilon_1 \in (0, \frac{1}{2}\epsilon_0]$ such that

$$\|\boldsymbol{x}^K - \tilde{\boldsymbol{x}}^*\| \leq \epsilon_0,$$

whenever $\|\boldsymbol{x}^{K-1} - \tilde{\boldsymbol{x}}^*\| \leq \epsilon_1$. At first, we have $T_\gamma^t(\boldsymbol{x}^{K-1}) = \overline{T}_\gamma^t(\boldsymbol{x}^{K-1})$ since $\boldsymbol{x}^{K-1}$ is the interior point and Proposition 25, we have

$$\begin{aligned}\|\boldsymbol{x}^K - \tilde{\boldsymbol{x}}^*\| &= \|T_\gamma^t(\boldsymbol{x}^{K-1}) - \overline{T}_\gamma^t(\boldsymbol{x}^{K-1}) + \overline{T}_\gamma^t(\boldsymbol{x}^{K-1}) - \overline{T}_\gamma^t(\tilde{\boldsymbol{x}}^*) + \overline{T}_\gamma^t(\tilde{\boldsymbol{x}}^*) - \tilde{\boldsymbol{x}}^*\|_2 \\ &= \|\overline{T}_\gamma^t(\boldsymbol{x}^{K-1}) - \overline{T}_\gamma^t(\tilde{\boldsymbol{x}}^*)\|_2.\end{aligned}$$

Moreover, due to Theorem 20, we know the mapping $\overline{T}_\gamma^t$ is continuous. Thus, for any $\epsilon_0 > 0$, there exists some constants $\delta > 0$ such that $\|\boldsymbol{x}^{K-1} - \tilde{\boldsymbol{x}}^*\|_2 < \delta$, we have $\|\overline{T}_\gamma^t(\boldsymbol{x}^{K-1}) - \overline{T}_\gamma^t(\tilde{\boldsymbol{x}}^*)\|_2 < \epsilon_0$. We can always choose a small constant $\epsilon' < \min\{\delta, \frac{1}{2}\epsilon_0\}$ to make the above argument hold.

Repeating the above argument K times, there exists a sequence $\{\epsilon_k\}_{k=0}^K$ such that $\epsilon_{k+1} \leq \frac{1}{2}\epsilon_k$ and

$$\|\boldsymbol{x}^{K-k} - \tilde{\boldsymbol{x}}^*\| \leq \epsilon_k,$$

whenever we have $\|\boldsymbol{x}^{K-k-1} - \tilde{\boldsymbol{x}}^*\| \leq \epsilon_{k+1}$, for any $k \in [K]$.

Now, we are ready to construct the initial point, i.e., $\boldsymbol{x}^0 \in \mathbb{B}_{\epsilon_K}(\tilde{\boldsymbol{x}}^*) \cap \mathcal{X} \cap \operatorname{int}(\operatorname{dom}(h))$. We set $\epsilon_0 = \epsilon$ and get

$$\boldsymbol{x}^k \in \mathbb{B}_{\epsilon_{K-k}}(\tilde{\boldsymbol{x}}^*) \subseteq \mathbb{B}_\epsilon(\tilde{\boldsymbol{x}}^*) \text{ for } k = 0, 1, \ldots, K.$$

We complete our proof. ■

E.2. Proof of Proposition 9

Proof Since $\operatorname{cl}(\operatorname{dom}(h)) = \mathbb{R}^n_+$, then $\operatorname{cl}(\operatorname{dom}(\varphi)) = \mathbb{R}_+$, $\mathcal{I}(\boldsymbol{x}) = \{i \in [n] : x_i > 0\}$, and $\mathcal{B}(\boldsymbol{x}) = \{b \in [n] : x_b = 0\}$. Moreover, we have $g(\boldsymbol{x}) = \delta_{\mathcal{X}}(\boldsymbol{x})$ and

$$\partial g(\boldsymbol{x}) = \{A^T\boldsymbol{\mu} - \boldsymbol{\lambda} : \boldsymbol{\mu} \in \mathbb{R}^m, \lambda_i = 0\ \forall i \in \mathcal{I}(\boldsymbol{x}), \lambda_b \geq 0\ \forall b \in \mathcal{B}(\boldsymbol{x})\}.$$

The compactness of $\mathcal{X}$ ensures the existence of $\tilde{\boldsymbol{x}}^* = \operatorname{argmax}_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$. Since $\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$ is convex problem and f is not a constant on $\mathcal{X}$, we have $f(\tilde{\boldsymbol{x}}^*) \neq \min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$. Thus, from optimality condition, we have

$$0 \notin \nabla f(\tilde{\boldsymbol{x}}^*) + \partial g(\tilde{\boldsymbol{x}}^*). \tag{E.1}$$

By contrast, $\tilde{\boldsymbol{x}}^*$ is the optimal solution of problem $\min_{\boldsymbol{x}\in\mathcal{X}} -f(\boldsymbol{x})$, whose optimality condition yields $\mathbf{0} \in -\nabla f(\tilde{\boldsymbol{x}}^*) + \partial\delta_{\mathcal{X}}(\tilde{\boldsymbol{x}}^*)$, which is equivalent to

$$\mathbf{0} \in \{-\nabla f(\tilde{\boldsymbol{x}}^*) + A^T\boldsymbol{\mu} - \boldsymbol{\lambda} : \boldsymbol{\mu} \in \mathbb{R}^m, \lambda_i = 0\ \forall i \in \mathcal{I}(\tilde{\boldsymbol{x}}^*), \lambda_b \geq 0\ \forall b \in \mathcal{B}(\tilde{\boldsymbol{x}}^*)\}$$

It follows that there exists $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ such that

$$\boldsymbol{\lambda}_{\mathcal{B}(\tilde{\boldsymbol{x}}^*)} \geq \mathbf{0},\ \boldsymbol{\lambda}_{\mathcal{I}(\tilde{\boldsymbol{x}}^*)} = \mathbf{0},\ \text{and } \mathbf{0} = -\nabla f(\tilde{\boldsymbol{x}}^*) + A^T\boldsymbol{\mu} - \boldsymbol{\lambda}.$$

Let $\boldsymbol{p} = -\boldsymbol{\lambda}$. Then, we have

$$\boldsymbol{p} = \nabla f(\tilde{\boldsymbol{x}}^*) - A^T\boldsymbol{\mu} \in \nabla f(\tilde{\boldsymbol{x}}^*) + \partial g(\tilde{\boldsymbol{x}}^*) \text{ and } \boldsymbol{p}_{\mathcal{I}(\tilde{\boldsymbol{x}}^*)} = 0.$$

Together with (E.1), we can conclude that $\tilde{\boldsymbol{x}}^*$ is a spurious point. ■

Appendix F. Examples of spurious stationary points

We present simple counter-examples, both convex and non-convex, to illustrate the presence of spurious stationary points.

Example 3 (Convex counter-example) *Suppose that* $\mathrm{cl}(\mathrm{dom}(h)) = \mathbb{R}^2_+$ *and consider the following simple problem:*

$$\begin{aligned} \min_{x_1,x_2} \quad & -x_1 \\ \text{s.t.} \quad & x_1 + x_2 = 1, x_1, x_2 \geq 0. \end{aligned}$$

The point $(0,1)$ *is identified as a spurious stationary point. We can determine the interior coordinate* $\mathcal{I}((0,1)) = 2$ *and compute the subdifferential at the point* $(0,1)$ *as*

$$\begin{aligned} \partial F((0,1)) &= (-1,0) + \mathcal{N}_{\{\boldsymbol{x}\in\mathbb{R}^2_+ : x_1+x_2=1\}}((0,1)) \\ &= \{(-1,0) + \lambda(-1,0) + \mu(1,1) : \lambda \in \mathbb{R}_+, \mu \in \mathbb{R}\}. \end{aligned}$$

Consequently, we find that $\mathbf{0} \notin \partial F((0,1))$ *and* $\boldsymbol{p} = (-1,0) \in \partial F((0,1))$ *with* $\boldsymbol{p}_{\mathcal{I}((0,1))} = p_2 = 0$.

Remark 30 *It is worth noting that in Example 3, for all feasible points* $\boldsymbol{x}$ *lying in the interior,*

$$\mathrm{dist}(\mathbf{0}, \partial F(\boldsymbol{x})) = \min_{\mu\in\mathbb{R}} \|(-1,0) + \mu(1,1)\| = \frac{\sqrt{2}}{2}.$$

Since a sequence $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$ *generated by the BPs belong to the interior of the kernel domain, we see that* $\mathrm{dist}(\mathbf{0}, \partial F(\boldsymbol{x}^k)) \equiv 0$ *for all* $k \in \mathbb{N}$. *Hence, the minimal subdifferential norm* $\mathrm{dist}(\mathbf{0}, \partial F)$ *is not a suitable measure for the BPs.*

Example 4 (Nonconvex counter-example) *Suppose that* $\mathrm{cl}(\mathrm{dom}(h)) = \mathbb{R}^2_+$ *and consider the following simple problem:*

$$\begin{aligned} \min_{x_1,x_2} \quad & f(x_1,x_2) = -x_1^2 + x_2 \\ \text{s.t.} \quad & x_1 + x_2 = 1, x_1, x_2 \geq 0. \end{aligned}$$

Similar with the convex case, the point $(0,1)$ *is identified as a spurious stationary point. We can determine that the interior coordinate* $\mathcal{I}((0,1)) = 2$ *and compute the subdifferential at the point* $(0,1)$ *as*

$$\partial F((0,1)) = \{(0,1) + \lambda(-1,0) + \mu(1,1) : \lambda \in \mathbb{R}_+, \mu \in \mathbb{R}\}.$$

Consequently, we find that $\mathbf{0} \notin \partial F((0,1))$ *and* $\boldsymbol{p} = (-1,0) \in \partial F((0,1))$ *with* $\boldsymbol{p}_{\mathcal{I}((0,1))} = p_2 = 0$.

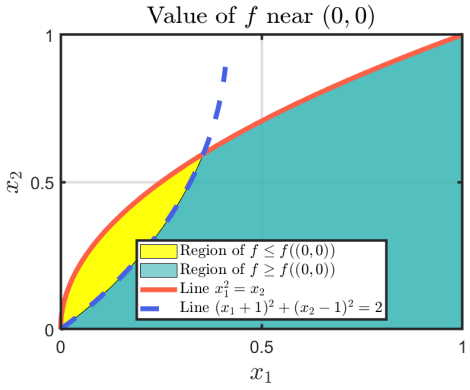

Figure 1: Illustration for Example 5

One may wonder whether spurious stationary points must be (locally) maximal points. The following example reveals that this is not true.

Example 5 *Consider the convex problem*

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^2_+} \quad & f(x_1,x_2) = (x_1+1)^2 + (x_2-1)^2 \\ \text{s.t.} \quad & x_2^2 \le x_1. \end{aligned}$$

The point $(0,0)$ *is identified as a spurious stationary point. We can determine the interior coordinate* $\mathcal{I}((0,0)) = \varnothing$ *and compute the subdifferential at the point* $(0,0)$ *as*

$$\begin{aligned} \partial F((0,0)) &= (2,-2) + \mathcal{N}_{\{\boldsymbol{x}\in\mathbb{R}^2_+ : x_2^2 \le x_1\}}((0,0)) \\ &= \left\{(2,-2) + \lambda_1(-1,0) + \lambda_2(-1,0) + \lambda_3(0,-1) : \boldsymbol{\lambda} \in \mathbb{R}^3_+\right\}. \end{aligned}$$

Consequently, we find that $\mathbf{0} \notin \partial F((0,0))$*, and hence* $(0,0)$ *is a spurious stationary point.*

In Example 5, the spurious stationary point $(0,0)$ is not a (local) maximal point. We illustrate this fact through Figure 1. Clearly, the region where $F \le F((0,0))$, i.e., the yellow part that satisfies $(x_1+1)^2 + (x_2-1)^2 \le 2, x_2^2 \le x_1$, and $\boldsymbol{x} \in \mathbb{R}^2_+$, is non-empty.

Finally, to enhance our understanding of the finite step trap behavior nearby spurious points, we give several instances on escaping from spurious stationary points.

Example 6 *We revisited the convex counter-example presented in Example 3, where the unique spurious point is* $\tilde{\boldsymbol{x}}^* = (0,1)$*. We choose the kernel function as the negative entropy* $\varphi(x) = x\log(x)$*, a popular choice for managing simplex constraints. For any* $K \in \mathbb{N}$ *and* $\epsilon > 0$*, we construct the initial point* $\boldsymbol{x}^0$ *as follows:*

$$\boldsymbol{x}^0 = \left(\frac{\sqrt{2}\epsilon}{2}e^{-tK}, 1 - \frac{\sqrt{2}\epsilon}{2}e^{-tK}\right).$$

Moreover, for all $k \in [K]$*, we apply the standard Bregman gradient descent method as*

$$\boldsymbol{x}^{k+1} = \arg\min_{\boldsymbol{y}} t(-1,0)^T\boldsymbol{y} + D_h(\boldsymbol{y},\boldsymbol{x}^k) + \delta_{\Delta_2}(\boldsymbol{y})$$

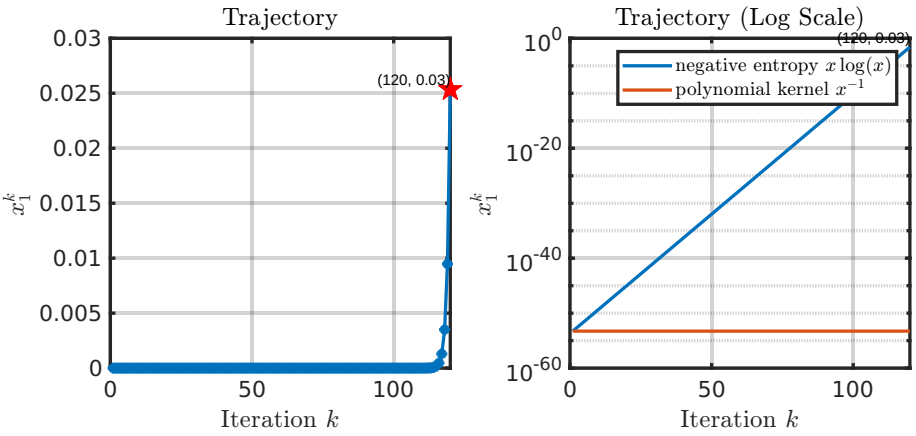

Figure 2: The trajectory plot for $\{x_1^k\}_{k\in[K]}$ in Example 6 and 7 ($\alpha = 1$), with $K = 120$ and $\epsilon = 0.1$. The initial point is chosen according to the negative entropy kernel scenario.

$$= \left(\frac{x_1^k}{x_1^k + e^{-t}x_2^k}, \frac{e^{-t}x_2^k}{x_1^k + e^{-t}x_2^k} \right), \quad \forall k \in [K].$$

Then, we can quantify the distance between $\boldsymbol{x}^{k+1}$ and $\tilde{\boldsymbol{x}}^*$ as

$$\|\boldsymbol{x}^{k+1} - \tilde{\boldsymbol{x}}^*\| = \frac{\sqrt{2}x_1^k}{x_1^k + e^{-t}x_2^k} \leq \sqrt{2}e^t x_1^k \leq \sqrt{2}e^{tk}x_1^0 = e^{-t(K-k)}\epsilon \leq \epsilon,$$

where the first inequality is derived from the constraint $x_1^k + x_2^k = 1$ and $t \geq 0$, the second inequality is justified by iteratively applying the recursive relation from the first inequality k times.

From the example involving negative entropy, it becomes clear that constructing the initial point at a distance that exponentially decays with respect to the spurious point is crucial. Later on, we want to provide another artificially constructed example to demonstrate the importance of the kernel function's growth condition in determining the necessary distance between the initial point and the spurious point to trigger a finite step trap. Essentially, the challenge of falling into a finite step trap varies significantly across different kernel functions.

Example 7 (Polynomial kernel) *We still consider the convex counter-example presented in Example 3 with a kernel function as $\varphi(x) = \frac{1}{\alpha}x^{-\alpha}$ where $\alpha > 0$. For any $K \in \mathbb{N}$ and $\epsilon > 0$, we construct the initial point $\boldsymbol{x}^0$ as follows:*

$$\boldsymbol{x}_1^0 = \min\left\{ \left(\frac{2}{tK + \epsilon^{-(\alpha+1)}} \right)^{\frac{1}{\alpha+1}}, \frac{1}{1+2^{\alpha+1}} \right\}$$

and $x_2^0 = 1 - x_1^0$. Moreover, for all $k \in [K]$, we apply the standard Bregman gradient descent method as

$$\boldsymbol{x}^{k+1} = \arg\min_{\boldsymbol{y}} t(-1,0)^T\boldsymbol{y} + D_h(\boldsymbol{y}, \boldsymbol{x}^k) + \delta_{\Delta_2}(\boldsymbol{y}).$$

If $\boldsymbol{x}^k \in \operatorname{int}(\operatorname{dom}(h)) \cap \mathcal{X}$, we have $\boldsymbol{x}^{k+1} \in \operatorname{int}(\operatorname{dom}(h)) \cap \mathcal{X}$ from Lemma 2. We can write down its optimality condition:

$$\begin{cases} -t + \varphi'(x_1^{k+1}) - \varphi'(x_1^k) + \mu_{k+1} = 0 \\ \varphi'(x_2^{k+1}) - \varphi'(x_2^k) + \mu_{k+1} = 0 \\ x_1^{k+1} + x_2^{k+1} = 1. \end{cases}$$

Summing up the above equation from $k = 0$ to $k = K - 1$, we have

$$\begin{cases} -tK + \varphi'(x_1^K) - \varphi'(x_1^0) + \sum_{k=1}^{K} \mu_{k+1} = 0 \\ \varphi'(x_2^K) - \varphi'(x_2^0) + \sum_{k=1}^{K} \mu_{k+1} = 0. \end{cases}$$

Since $\varphi'(x) = -x^{-\alpha-1}$ and $(\varphi')^{-1}(y) = \left(-\frac{1}{y}\right)^{\frac{1}{1+\alpha}}$, we know φ' is negative on $\mathbb{R}_{++}$ and $(\varphi')^{-1}$ is monotonically increasing on $\mathbb{R}_{-}$. Then, we have

$$-\sum_{k=1}^{K} \mu_{k+1} = \varphi'(x_2^K) - \varphi'(x_2^0) \leq -\varphi'(x_2^0).$$

Finally, we proceed to bound x_1^K

$$\begin{aligned} x_1^K &= (\varphi')^{-1}\left(\varphi'(x_1^0) - \sum_{k=1}^{K} \mu_{k+1} + tK\right) \\ &\leq (\varphi')^{-1}\left(\varphi'(x_1^0) - \varphi'(1 - x_1^0) + tK\right) \\ &= \left((x_1^0)^{-\alpha-1} - (1 - x_1^0)^{-\alpha-1} - tK\right)^{\frac{-1}{1+\alpha}} \leq \epsilon. \end{aligned}$$

where the first inequality follows from $-\sum_{k=1}^{K} \mu_{k+1} \leq -\varphi'(x_2^0)$ and $x_1^0 + x_2^0 = 1$ and the last inequality follows from

$$(x_1^0)^{-\alpha-1} - (1 - x_1^0)^{-\alpha-1} \geq (2x_1^0)^{-\alpha-1}$$

due to $x_1^0 \leq \frac{1}{1+2^{\alpha+1}}$ and $(2x_1^0)^{-\alpha-1} - tK > 0$. For simplicity, we ignore the final constant $\sqrt{2}$.

Remark 31 *From Figure 2, it is evident that despite both kernels facing the challenges outlined in Theorem 7, the nonnegative entropy kernel demonstrates superior performance compared to the polynomial kernel. This advantage can be attributed to the curvature information encapsulated by the inverse mapping of $\nabla h(\boldsymbol{x})$. Specifically, when $h(\boldsymbol{x}) = \boldsymbol{x}\log(\boldsymbol{x})$, $\nabla^{-1}h(\boldsymbol{x})$ exhibits exponential growth behavior, enabling each iteration to escape the unfavorable point by at least doubling the distance from it.*