Spurious Stationarity and Hardness Results for Bregman Proximal-type Algorithms

Jiajin Li

Sauder School of Business

University of British Columbia

UBC SAUDER SCHOOL OF BUSINESS

September 21, 2024

Joint work with He Chen (CUHK) and Anthony Man-Cho So (CUHK).

Main Results

1. Spurious stationary points inevitably exist when non-gradient Lipschitz kernels are used for Bregman proximal-type algorithms.

2. (Algorithm-Dependent Hardness Results) Bregman proximal-type algorithms *are unable to* escape from a spurious stationary point in finite steps when the initial point is bad.

Main Results

1. Spurious stationary points inevitably exist when non-gradient Lipschitz kernels are used for Bregman proximal-type algorithms.

2. (Algorithm-Dependent Hardness Results) Bregman proximal-type algorithms *are unable to* escape from a spurious stationary point in finite steps when the initial point is bad.

What we'll cover today?

1. **Introduction and Problem Settings**
2. Spurious Stationary Points and Examples
3. Algorithm-Dependent Hardness Results and their Implications
4. Unsatisfactory Stationary Measures and Convergence Behaviour Investigation

Mirror Descent (Non-Euclidean Gradient Descent)

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} F(\boldsymbol{x}) \tag{P}$$

- Gradient Descent:

$$\begin{aligned}\boldsymbol{x}_+ &= \boldsymbol{x} - t \cdot \nabla F(\boldsymbol{x}) \\ &= \arg\min_{\boldsymbol{y} \in \mathbb{R}^n} \nabla F(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2t} \|\boldsymbol{y} - \boldsymbol{x}\|^2.\end{aligned}$$

- Mirror Descent:

$$\boldsymbol{x}_+ = \arg\min_{\boldsymbol{y} \in \mathbb{R}^n} \nabla F(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2t} \underbrace{D_h(\boldsymbol{y}, \boldsymbol{x})}_{\text{Bregman Divergence}}.$$

Better to exploit the geometry of the problem at hand!

Mirror Descent (Non-Euclidean Gradient Descent)

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} F(\boldsymbol{x}) \tag{P}$$

- Gradient Descent:

$$\begin{aligned}\boldsymbol{x}_+ &= \boldsymbol{x} - t\cdot\nabla F(\boldsymbol{x})\\ &= \arg\min_{\boldsymbol{y}\in\mathbb{R}^n} \nabla F(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x}) + \frac{1}{2t}\|\boldsymbol{y}-\boldsymbol{x}\|^2.\end{aligned}$$

- Mirror Descent:

$$\boldsymbol{x}_+ = \arg\min_{\boldsymbol{y}\in\mathbb{R}^n} \nabla F(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x}) + \frac{1}{2t}\underbrace{D_h(\boldsymbol{y},\boldsymbol{x})}_{\textbf{Bregman Divergence}}.$$

Better to exploit the geometry of the problem at hand!

Mirror Descent (Non-Euclidean Gradient Descent)

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} F(\boldsymbol{x}) \tag{P}$$

- Gradient Descent:

$$\begin{aligned} \boldsymbol{x}_+ &= \boldsymbol{x} - t \cdot \nabla F(\boldsymbol{x}) \\ &= \arg\min_{\boldsymbol{y} \in \mathbb{R}^n} \nabla F(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2t} \|\boldsymbol{y} - \boldsymbol{x}\|^2. \end{aligned}$$

- Mirror Descent:

$$\boldsymbol{x}_+ = \arg\min_{\boldsymbol{y} \in \mathbb{R}^n} \nabla F(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2t} \underbrace{D_h(\boldsymbol{y}, \boldsymbol{x})}_{\textbf{Bregman Divergence}} .$$

Better to exploit the geometry of the problem at hand!

Bregman Divergence

Definition

The Bregman divergence between two points $\boldsymbol{x}, \boldsymbol{y}$ associated with a kernel function $h : \Omega \to \mathbb{R}$ is defined as

$$D_h(\boldsymbol{x}, \boldsymbol{y}) := h(\boldsymbol{x}) - h(\boldsymbol{y}) - \nabla h(\boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y}),$$

where h is continuously differentiable and strictly convex on the convex set Ω.

- (✓) If $h(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|^2$, we have $D_h(\boldsymbol{y}, \boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2$. Then, mirror descent reduces to the vanilla gradient descent.
- (✗) If $h(\boldsymbol{x}) = \sum_{i=1}^n x_i \log x_i$, $D_h(\boldsymbol{y}, \boldsymbol{x})$ is just KL divergence.
- (✗) Other non-gradient Lipschitz kernel functions $\cdots$

Optimization Problems

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

- $\mathrm{dom}(g) = \mathcal{X}$ is a nonempty closed convex set.
- $f : \mathcal{X} \to \mathbb{R}$ is continuous differentiable on $\mathcal{X}$ (possibly **nonconvex**).
- $g : \mathcal{X} \to \overline{\mathbb{R}}$ is **convex** and **locally Lipschitz continuous**, e.g.,
    - Indicator function -> include the constrained optimization problem as a special case.

Bregman Proximal-Type Algorithms

$$x_+ = T_\gamma^t(x) := \arg\min_{y \in \mathbb{R}^n} \left\{ \underbrace{\gamma(y;x)}_{\textbf{Surrogate Model}} + g(y) + \frac{1}{t} D_h(y,x) \right\} \quad (\mathcal{A})$$

- If $\gamma(\boldsymbol{y};\boldsymbol{x}) = f(\boldsymbol{y})$, $(\mathcal{A})$ reduces to Bregman proximal point methods [Chen and Teboulle, 1993].
- If $\gamma(\boldsymbol{y};\boldsymbol{x}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x})$, $(\mathcal{A})$ reduces to Bregman proximal (projected) gradient descent [Bauschke et al., 2017],[Bauschke et al., 2019].
- If $\gamma(\boldsymbol{y};\boldsymbol{x}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{x})^T\nabla^2 f(\boldsymbol{x})(\boldsymbol{y}-\boldsymbol{x})$, $(\mathcal{A})$ has been recently explored by [Doikov and Nesterov, 2023].

Separable Kernel Functions

- $h(\boldsymbol{x}) = \sum_{i=1}^{n} \varphi(x_i)$, where $\varphi : \mathbb{R} \to \overline{\mathbb{R}}$ is a univariate function.
- φ is continuously differentiable on $\mathrm{int}(\mathrm{dom}(\varphi))$, and $|\varphi'(x^k)| \to +\infty$ as $x^k \to x \in \mathrm{bd}(\mathrm{dom}(\varphi))$.
- φ is strictly convex.

1. Boltzmann–Shannon entropy kernel $h(\boldsymbol{x}) = \sum_{i=1}^{n} x_i \log(x_i)$;
2. Fermi–Dirac entropy kernel $h(\boldsymbol{x}) = \sum_{i=1}^{n} x_i \log(x_i) + (1 - x_i) \log(1 - x_i)$;
3. Burg entropy kernel $h(\boldsymbol{x}) = \sum_{i=1}^{n} - \log(x_i)$;
4. Fractional power kernel $h(\boldsymbol{x}) = \sum_{i=1}^{n} px_i - \frac{x_i^p}{1-p}$ $(0 < p < 1)$;
5. Hellinger entropy kernel $h(\boldsymbol{x}) = \sum_{i=1}^{n} -\sqrt{1 - x_i^2}$.

Separable Kernel Functions

- $h(\boldsymbol{x}) = \sum_{i=1}^{n} \varphi(x_i)$, where $\varphi : \mathbb{R} \to \overline{\mathbb{R}}$ is a univariate function.
- φ is continuously differentiable on $\mathrm{int}(\mathrm{dom}(\varphi))$, and $|\varphi'(x^k)| \to +\infty$ as $x^k \to x \in \mathrm{bd}(\mathrm{dom}(\varphi))$.
- φ is strictly convex.

1. Boltzmann–Shannon entropy kernel $h(\boldsymbol{x}) = \sum_{i=1}^{n} x_i \log(x_i)$;
2. Fermi–Dirac entropy kernel $h(\boldsymbol{x}) = \sum_{i=1}^{n} x_i \log(x_i) + (1 - x_i) \log(1 - x_i)$;
3. Burg entropy kernel $h(\boldsymbol{x}) = \sum_{i=1}^{n} - \log(x_i)$;
4. Fractional power kernel $h(\boldsymbol{x}) = \sum_{i=1}^{n} px_i - \frac{x_i^p}{1-p}$ $(0 < p < 1)$;
5. Hellinger entropy kernel $h(\boldsymbol{x}) = \sum_{i=1}^{n} -\sqrt{1 - x_i^2}$.

What we'll cover today?

1. Introduction and Problem Settings
2. **Spurious Stationary Points and Examples**
3. Algorithm-Dependent Hardness Results and their Implications
4. Unsatisfactory Stationary Measures and Convergence Behaviour Investigation

Spurious Stationarity

Definition

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 $0 \notin \partial F(\boldsymbol{x})$.

- $\mathcal{I}(\boldsymbol{x}) := \{i \in [n] : x_i \in \mathrm{int}(\mathrm{dom}(\varphi))\}$.
- Spurious stationary points exist only when the kernel is **non-gradient Lipschitz**.
- For a kernel h with gradient Lipschitz property, we have $\mathrm{dom}(\varphi) = \mathbb{R}$ and $\mathcal{I}(\boldsymbol{x}) = [n]$ hold for all $\boldsymbol{x} \in \mathcal{X}$, thereby precluding the existence of spurious stationary points.

Only depends on the problem itself and the kernel function!

Spurious Stationarity

Definition

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 $0 \notin \partial F(\boldsymbol{x})$.

- $\mathcal{I}(\boldsymbol{x}) := \{i \in [n] : x_i \in \mathrm{int}(\mathrm{dom}(\varphi))\}$.
- Spurious stationary points exist only when the kernel is **non-gradient Lipschitz**.
- For a kernel h with gradient Lipschitz property, we have $\mathrm{dom}(\varphi) = \mathbb{R}$ and $\mathcal{I}(\boldsymbol{x}) = [n]$ hold for all $\boldsymbol{x} \in \mathcal{X}$, thereby precluding the existence of spurious stationary points.

Only depends on the problem itself and the kernel function!

Spurious Stationarity

Definition

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 $0 \notin \partial F(\boldsymbol{x})$.

- $\mathcal{I}(\boldsymbol{x}) := \{i \in [n] : x_i \in \mathrm{int}(\mathrm{dom}(\varphi))\}$.
- Spurious stationary points exist only when the kernel is **non-gradient Lipschitz**.
- For a kernel h with gradient Lipschitz property, we have $\mathrm{dom}(\varphi) = \mathbb{R}$ and $\mathcal{I}(\boldsymbol{x}) = [n]$ hold for all $\boldsymbol{x} \in \mathcal{X}$, thereby precluding the existence of spurious stationary points.

Only depends on the problem itself and the kernel function!

Convex Example

Example (A Simple Linear Programming Problem)

Suppose that $\mathrm{cl}(\mathrm{dom}(h)) = \mathbb{R}^2_+$ and consider the following simple problem:

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

The point (0,1) is identified as a spurious stationary point.

We find that $0 \notin \partial F((0,1))$ and $\boldsymbol{p} = (-1, 0) \in \partial F((0,1))$ with $\boldsymbol{p}_{\mathcal{I}((0,1))} = p_2 = 0$, i.e.,

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

Convex Example

Example (A Simple Linear Programming Problem)

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

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

Nonconvex Example

Example

Suppose that $\mathrm{cl}(\mathrm{dom}(h)) = \mathbb{R}^2_+$ and consider the following simple problem:

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

The point (0,1) is identified as a spurious stationary point.

Proposition (Existence of Spurious Stationary Points)

Consider a convex optimization problem

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathcal{X}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & A\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geq 0. \end{aligned}$$

Suppose the constraint set is compact and and f is non-constant. If $\mathrm{cl}(\mathrm{dom}(h)) = \mathbb{R}^n_+$, then **every maximal point** *$\tilde{\boldsymbol{x}}^* \in \mathrm{argmax}_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$ is a spurious stationary point.*

Nonconvex Example

Example

Suppose that $\mathrm{cl}(\mathrm{dom}(h)) = \mathbb{R}^2_+$ and consider the following simple problem:

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

The point (0,1) is identified as a spurious stationary point.

Proposition (Existence of Spurious Stationary Points)

Consider a convex optimization problem

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathcal{X}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & A\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geq 0. \end{aligned}$$

Suppose the constraint set is compact and and f is non-constant. If $\mathrm{cl}(\mathrm{dom}(h)) = \mathbb{R}^n_+$, then **every maximal point** $\tilde{\boldsymbol{x}}^* \in \mathrm{argmax}_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$ *is a spurious stationary point.*

What we'll cover today?

1. Introduction and Problem Settings
2. Spurious Stationary Points and Examples
3. **Algorithm-Dependent Hardness Results and their Implications**
4. Unsatisfactory Stationary Measures and Convergence Behaviour Investigation

Algorithm-dependent Hardness Result

Theorem

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* $\boldsymbol{x}^0 \in \mathcal{B}_\epsilon(\tilde{\boldsymbol{x}}^*) \cap \mathcal{X}$*, sufficiently close to the* **spurious stationary point** $\tilde{\boldsymbol{x}}^*$*, such that*

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

[Corollary 1, Bauschke et al., 2017]. When f is convex, the sequence $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$ generated by BPG satisfies

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

where $\overline{\boldsymbol{x}} \in \mathrm{argmin}_{x\in\mathcal{X}} f$ is the global minimizer, t is the step size, and $\boldsymbol{x}^0$ is an arbitrary initial point.

Algorithm-dependent Hardness Result

Theorem

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* $\boldsymbol{x}^0 \in \mathcal{B}_\epsilon(\tilde{\boldsymbol{x}}^*) \cap \mathcal{X}$*, sufficiently close to the* **spurious stationary point** $\tilde{\boldsymbol{x}}^*$*, such that*

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

[Corollary 1, Bauschke et al., 2017]. When f is convex, the sequence $\{\boldsymbol{x}^k\}_{k\in\mathbb{N}}$ generated by BPG satisfies

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

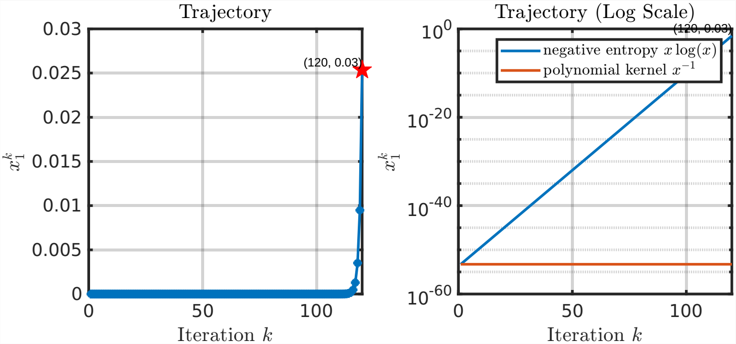
where $\overline{\boldsymbol{x}} \in \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{X}} f$ is the global minimizer, t is the step size, and $\boldsymbol{x}^0$ is an arbitrary initial point.

Visualization

Example (The Simple Linear Programming Problem)

For every K and $\epsilon > 0$, we construct the initial point as

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



What we'll cover today?

1. Introduction and Problem Settings
2. Spurious Stationary Points and Examples
3. Algorithm-Dependent Hardness Results and their Implications
4. **Unsatisfactory Stationary Measures and Convergence Behaviour**

Residual Function

Understand how the sequence of iterations behaves and how close it gets to convergence.

A standard recipe in optimization:

- Propose a residual function $R : \mathbb{R}^n \to \mathbb{R}_+$ that measures the stationarity of the iterations.
- Estabilish the convergence of the sequence of $\{R(\boldsymbol{x}^k)\}_{k\geq 0}$.

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

Residual Function

Understand how the sequence of iterations behaves and how close it gets to convergence.

A standard recipe in optimization:

- Propose a residual function $R : \mathbb{R}^n \to \mathbb{R}_+$ that measures the stationarity of the iterations.
- Estabilish the convergence of the sequence of $\{R(\boldsymbol{x}^k)\}_{k\geq 0}$.

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

Residual Function

Understand how the sequence of iterations behaves and how close it gets to convergence.

A standard recipe in optimization:

- Propose a residual function $R : \mathbb{R}^n \to \mathbb{R}_+$ that measures the stationarity of the iterations.
- Estabilish the convergence of the sequence of $\{R(\boldsymbol{x}^k)\}_{k \geq 0}$.

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

Existing stationarity measures are not well-defined

All existing stationarity measure can be unified as

$$R_\gamma^t(\boldsymbol{x}) := D_h(T_\gamma^t(\boldsymbol{x}), \boldsymbol{x})$$

the relative change w.r.t Bregman divergence.

- R_γ^t is not well-defined on the boundary $\mathrm{bd}(\mathrm{dom}(h))$.
- The mapping $\boldsymbol{x} \mapsto T_\gamma^t(\boldsymbol{x})$ involves the Bregman divergence function $(\boldsymbol{y}, \boldsymbol{x}) \mapsto D_h(\boldsymbol{y}, \boldsymbol{x})$, which is only defined on $\mathrm{dom}(h) \times \mathrm{int}(\mathrm{dom}(h))$.

Existing stationarity measures are not well-defined

All existing stationarity measure can be unified as

$$R_\gamma^t(\boldsymbol{x}) := D_h(T_\gamma^t(\boldsymbol{x}), \boldsymbol{x})$$

the relative change w.r.t Bregman divergence.

- R_γ^t is not well-defined on the boundary $\mathrm{bd}(\mathrm{dom}(h))$.
- The mapping $\boldsymbol{x} \mapsto T_\gamma^t(\boldsymbol{x})$ involves the Bregman divergence function $(\boldsymbol{y}, \boldsymbol{x}) \mapsto D_h(\boldsymbol{y}, \boldsymbol{x})$, which is only defined on $\mathrm{dom}(h) \times \mathrm{int}(\mathrm{dom}(h))$.

Extended Stationarity Measure

A simple fix: Only account for the **interior coordinates**, i.e.,

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

where $\overline{T}_\gamma^t(\boldsymbol{x})$ denotes the update rule that ensures the boundary coordinates remain fixed.

$$\lim_{k\to\infty} R(\boldsymbol{x}^k) = 0 \overset{?}{\iff} 0 \in \partial F\left(\lim_{k\to\infty} \boldsymbol{x}^k\right)$$

- The residual function is **continuous** (✓).
- The residual function equals to zeros **if and only** if x is a stationary point.
    - If x is a stationary point, we have $\overline{R}_\gamma^t(x) = 0$ (✓).
    - $\Rightarrow$?

Extended Stationarity Measure

A simple fix: Only account for the **interior coordinates**, i.e.,

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

where $\overline{T}_\gamma^t(\boldsymbol{x})$ denotes the update rule that ensures the boundary coordinates remain fixed.

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

- The residual function is **continuous** (✓).
- The residual function equals to zeros **if and only** if x is a stationary point.
  - If x is a stationary point, we have $\overline{R}_\gamma^t(x) = 0$ (✓).
  - $\Rightarrow$?

Extended Stationarity Measure

A simple fix: Only account for the **interior coordinates**, i.e.,

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

where $\overline{T}_\gamma^t(\boldsymbol{x})$ denotes the update rule that ensures the boundary coordinates remain fixed.

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

- The residual function is **continuous** (✓).
- The residual function equals to zeros **if and only** if x is a stationary point.
  - If x is a stationary point, we have $\overline{R}_\gamma^t(x) = 0$ (✓).
  - $\Rightarrow$?

Extended Stationarity Measure

A simple fix: Only account for the **interior coordinates**, i.e.,

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

where $\overline{T}_{\gamma}^{t}(\boldsymbol{x})$ denotes the update rule that ensures the boundary coordinates remain fixed.

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

- The residual function is **continuous** (✓).
- The residual function equals to zeros **if and only** if x is a stationary point.
  - If $\boldsymbol{x}$ is a stationary point, we have $\overline{R}_{\gamma}^{t}(\boldsymbol{x}) = 0$ (✓).
  - $\Rightarrow$?

Extended Stationarity Measure

A simple fix: Only account for the **interior coordinates**, i.e.,

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

where $\overline{T}_\gamma^t(\boldsymbol{x})$ denotes the update rule that ensures the boundary coordinates remain fixed.

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

- The residual function is **continuous** (✓).
- The residual function equals to zeros **if and only** if x is a stationary point.
  - If $\boldsymbol{x}$ is a stationary point, we have $\overline{R}_\gamma^t(\boldsymbol{x}) = 0$ (✓).
  - $\Rightarrow$?

Unsatisfactory Stationarity Measure and Spurious Stationary Points

Proposition (Characterization of spurious stationary points)

A point $\boldsymbol{x} \in \mathcal{X}$ is a spurious stationary point if and only if

$$\overline{R}^t_\gamma(\boldsymbol{x}) = 0 \ \mathit{but}\ 0 \notin \partial F(\boldsymbol{x}).$$

- We have demonstrated that spurious stationary points are ubiquitous.
- Can we provide a satisfactory stationarity measure? -> New Bregman-type algorithms...

He Chen*, Jiajin Li*, Anthony Man-Cho So*. Spurious Stationarity and Hardness Results for Mirror Descent http://arxiv.org/abs/2404.08073

Unsatisfactory Stationarity Measure and Spurious Stationary Points

Proposition (Characterization of spurious stationary points)

A point $\boldsymbol{x} \in \mathcal{X}$ *is a spurious stationary point if and only if*

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

- We have demonstrated that spurious stationary points are ubiquitous.
- Can we provide a satisfactory stationarity measure? -> New Bregman-type algorithms...

He Chen*, Jiajin Li*, Anthony Man-Cho So*. Spurious Stationarity and Hardness Results for Mirror Descent http://arxiv.org/abs/2404.08073

Unsatisfactory Stationarity Measure and Spurious Stationary Points

Proposition (Characterization of spurious stationary points)

A point $\boldsymbol{x} \in \mathcal{X}$ is a spurious stationary point if and only if

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

- We have demonstrated that spurious stationary points are ubiquitous.
- Can we provide a satisfactory stationarity measure? -> New Bregman-type algorithms...

He Chen*, Jiajin Li*, Anthony Man-Cho So*. Spurious Stationarity and Hardness Results for Mirror Descent http://arxiv.org/abs/2404.08073

Unsatisfactory Stationarity Measure and Spurious Stationary Points

Proposition (Characterization of spurious stationary points)

A point $\boldsymbol{x} \in \mathcal{X}$ is a spurious stationary point if and only if

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

- We have demonstrated that spurious stationary points are ubiquitous.
- Can we provide a satisfactory stationarity measure? -> New Bregman-type algorithms...

He Chen*, Jiajin Li*, Anthony Man-Cho So*. Spurious Stationarity and Hardness Results for Mirror Descent http://arxiv.org/abs/2404.08073

Thank you for your listening!

Any questions?