Mirror Descent in Stochastic Approximation

Fatim Majumder — Mentor: Katie Keegan

March 25, 2025

Introduction

Context and Motivation

- **Mirror Descent** is a robust optimization algorithm, particularly effective in *stochastic approximation scenarios* where the problem's geometry is complex and only partially known.
- Traditional techniques like *gradient descent* may struggle with uncertainty and noise in estimating gradients and constraints.
- **Adaptability:** Mirror Descent adapts by using *stochastic gradients* for efficient navigation through uncertain and non-standard geometrical spaces.
- This presentation explores **algorithm tailoring** for optimization in stochastic environments, emphasizing *geometric adaptability*.

Bregman Divergences

Understanding Bregman Divergence

- **Convex Function Setting:** Let $\mathcal{X}$ be a closed convex set. Consider $f : \mathcal{X} \to \mathbb{R}$, a continuously differentiable and strictly convex function.
- **First-Order Approximation:** For any point $y \in \mathcal{X}$, the first-order approximation of f at y offers a linear estimate at another point $x \in \mathcal{X}$.
- **Quantifying Deviation:** Bregman divergence measures the deviation of f from its linear approximation at point y, highlighting the curvature of f around y.
- **Geometric Relationship:** It reveals the geometric relationship between a function and its convex conjugate, emphasizing the geometry of the function.

Visualizing Bregman Divergence

- Visual illustrations aid in understanding Bregman divergence.
- Below is a graphical example:

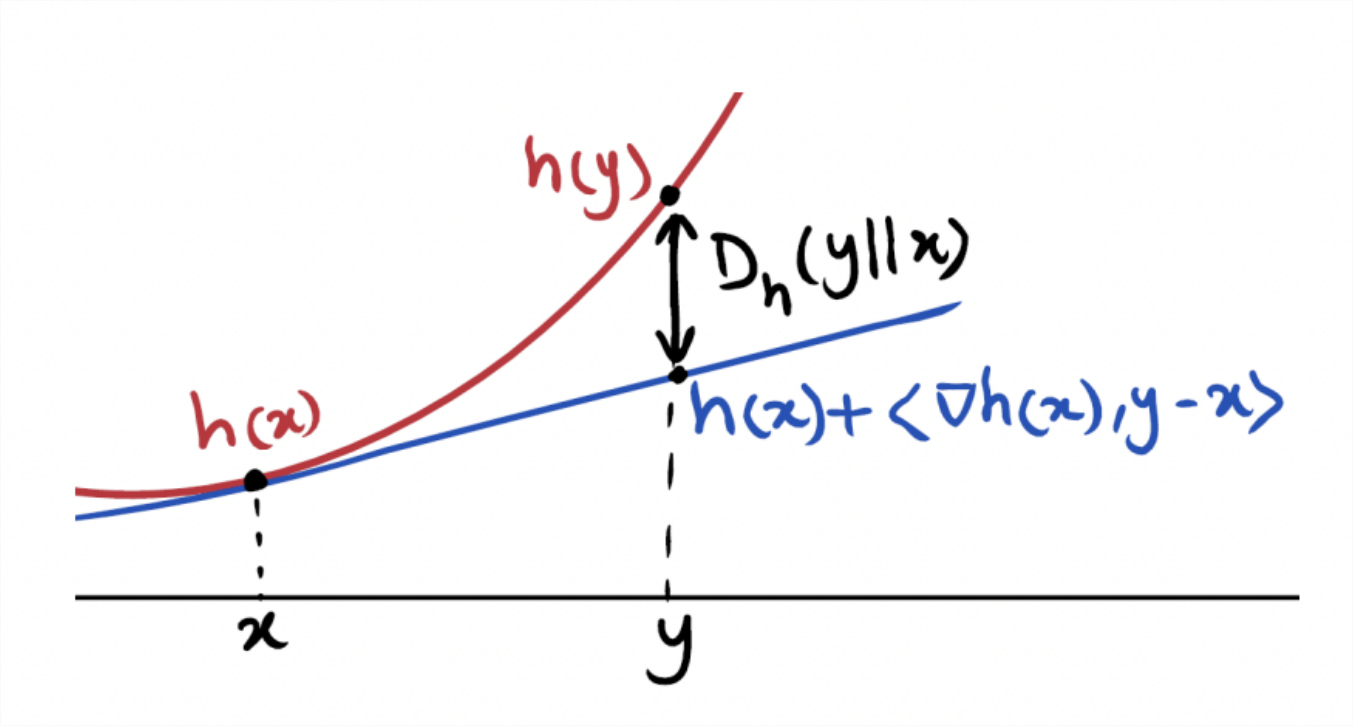

Figure 1: Graphical illustration of Bregman Divergence. Reference: CMU School of Computer Science

Formal Definition of Bregman Divergence

Definition
The Bregman divergence D_f for a function f is defined as:

$$D_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle,$$

where $\langle \cdot, \cdot \rangle$ represents the dot product. It quantifies the 'distance' between points x and y in the context of f.

- **Significance:** Bregman Divergence extends the concept of 'distance' beyond traditional Euclidean distance, adapting to the curvature of f.

Examples and Properties of Bregman Divergence

- **Dependence on Generating Function:** The form of Bregman divergence is dependent on the choice of the generating function h.
- **Kullback-Leibler Divergence:** When h is the negative entropy function ($h(x) = \sum x_i \log x_i$), the Bregman divergence becomes the Kullback-Leibler divergence, a key measure in information theory.

Key Properties of Bregman Divergence

- **Non-negativity:** $D_h(x, y) \geq 0$ for all x, y, indicating its role as a 'distance' measure.
- **Joint Convexity:** If h is strictly convex, then $D_h(x, y)$ is jointly convex in both x and y, an essential feature for optimization algorithms.

Mirror Descent: The Proximal Point View

Fundamental Property of Bregman Divergence

Context: For λ-strongly convex functions, Bregman divergence D_h quantifies discrepancy between $h(x)$ and its linear approximation at y.

Bregman Divergence Definition

$$D_h(x, y) = h(x) - \left[h(y) + \nabla h(y)^\top (x - y)\right],$$

Lower Bound

The divergence has a notable lower bound, indicating strong convexity:

$$D_h(x, y) \geq \frac{\lambda}{2}\|x - y\|^2 \geq 0.$$

Note: This property highlights the importance of Bregman divergence in optimization, especially with strongly convex functions.

The Mirror Map Φ in Mirror Descent

Context: The mirror descent framework relies on a well-defined mirror map Φ, which transforms the problem into a tractable form. Let $D \subseteq \mathbb{R}^n$ be an open domain and $\Phi : D \to \mathbb{R}$ be the designated mirror map.

$P1$: **Strict Convexity and Smoothness:** Φ should be strictly convex and continuously differentiable over D. This ensures a unique solution and a smooth mapping process.

$P2$: **Full Span in Dual Space:** The gradient of Φ should cover the entire dual space, i.e., $\{\nabla\Phi(x) : x \in D\} = \mathbb{R}^n$. This property guarantees that every direction in the dual space is reachable.

$P3$: **Unbounded Gradient at Boundary:** The gradient of Φ must become unbounded as it approaches the boundary of D: $\lim_{x\to\partial D} \|\nabla\Phi(x)\| = +\infty$.

Projection in Mirror Descent

Context: Bregman Projection in mirror descent ensures iterates remain within $\mathcal{X}$, a feasible set.

Definition
The Bregman Projection onto $\mathcal{X}$ minimizes Bregman divergence from y:

$$\Pi^{\mathcal{X}}(y) = \arg\min_{x \in \mathcal{X}} D_f(x, y).$$

- **Uniqueness:** Due to the strict convexity of D_f in x, the projection is unique.
- **Optimality Condition:** The process links the projected point to the gradients of f, guiding efficient optimization.

Mirror Descent through the Mirror Map

1. **Mapping to Dual Space:** Transform the current iterate x_t into the dual space using the mirror map Φ. This is achieved by computing $\theta_t = \nabla\Phi(x_t)$, where θ_t represents the dual coordinates of x_t.
2. **Gradient Step in Dual Space:** Perform a gradient descent step in the dual space. Update the dual variables by $\theta_{t+1} = \theta_t - \eta\nabla f(x_t)$, where η is the step size and $\nabla f(x_t)$ is the gradient of the objective function at x_t.
3. **Mapping Back to Primal Space:** Convert the updated dual variables back to the primal space to obtain the new iterate. This is done by computing $x'_{t+1} = \nabla\Phi^{-1}(\theta_{t+1})$.
4. **Projection onto Feasible Set:** Finally, project x'_{t+1} onto the feasible set $\mathcal{X}$ using the Bregman divergence to ensure the solution remains within the constraints of the problem.

Illustration of Mirror Descent Steps

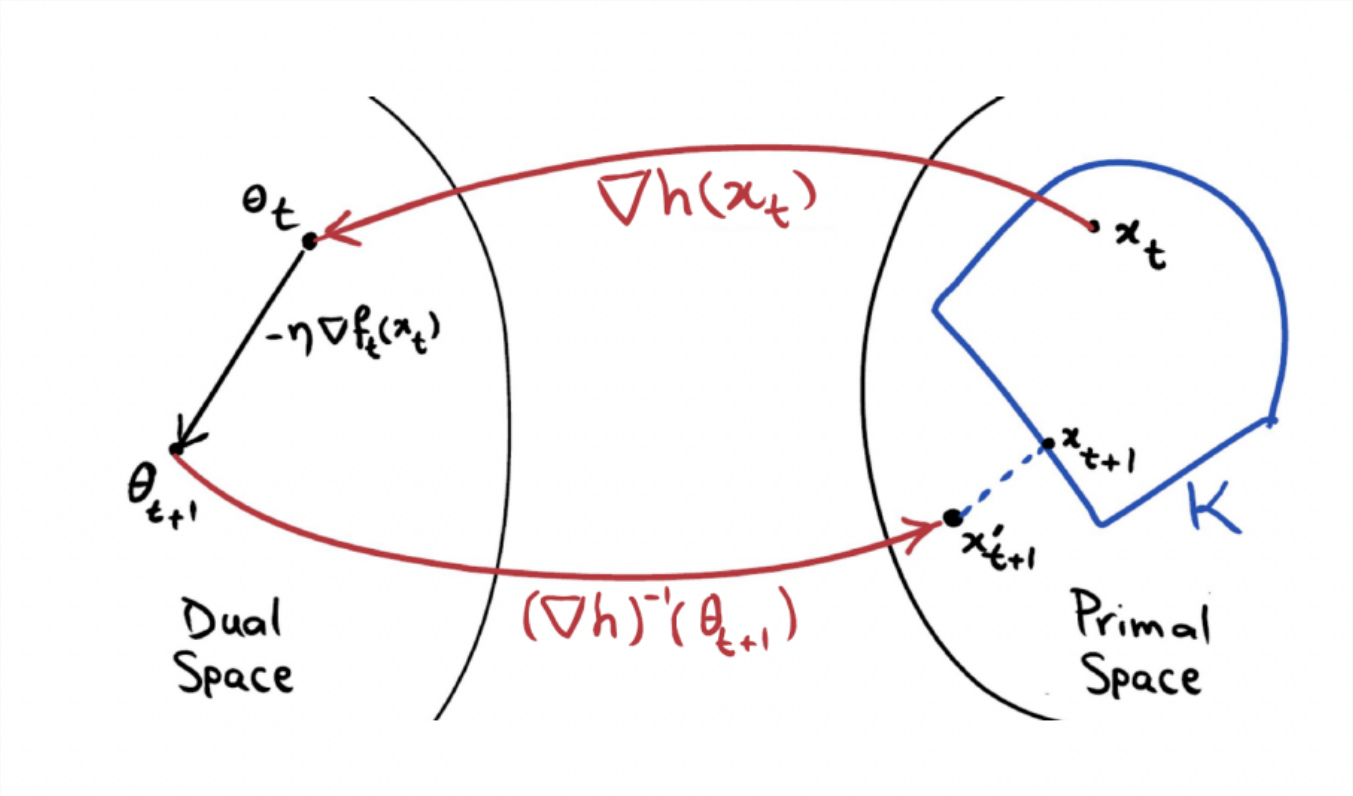

Figure 2: Iterative process of mirror descent visualized. Reference: CMU School of Computer Science

Constructing the Mirror Maps

- **Choose a Norm:** Select an appropriate norm for the problem.
- **Select a Function:** Identify a differentiable, strongly-convex function h compatible with the norm.
- **Example:** For $\|\cdot\|_2$, use $h(x) = \frac{1}{2}\|x\|_2^2$.
- **Construct Mirror Map:** Create the mirror map as the gradient of h: $\nabla h(x)$.
- **Application:** This mirror map is central to transformations in mirror descent.

The Mirror Descent Algorithm

Update Rule Formulation in Mirror Descent

Update Rule

The update process for obtaining the next iterate $x^{(k+1)}$ is formulated as an optimization problem:

$$x^{(k+1)} = \arg\min_x \left\{ f(x^{(k)}) + g^{(k)T}(x - x^{(k)}) + \frac{1}{\alpha_k} D_h(x, x^{(k)}) \right\},$$

where $D_h(x, y) = h(x) - \left[h(y) + \nabla h(y)^T (x - y) \right]$,

Here, $h(x)$ is a strongly convex function, and $D_h(x, y)$ represents the Bregman divergence between x and y.

Bregman Projection

The Bregman projection P_C^h, a key component of the update, ensures that $x^{(k+1)}$ stays within the feasible region C, maintaining the constraints of the problem.

Decomposing the Update Rule in Mirror Descent

Dual Space Optimization
The initial step is an optimization in the dual space, formulated as:

$$x^{(k+1)} = P_C^h \left(\arg\min_x \left\{ g^{(k)T} x + \frac{1}{\alpha_k} \left(h(x) - h(x^{(k)}) - \nabla h(x^{(k)})^T x \right) \right\} \right)$$

where $g^{(k)}$ is the gradient at iteration k, α_k is the step size, and h is the mirror map function.

Primal Space Mapping
After optimizing in the dual space, the solution is then mapped back to the primal space through Bregman projection:

$$P_C^h(y) = \arg\min_{x \in C} \left\{ h(x) - \nabla h(y)^T x \right\},$$

where C represents the feasible set.

Understanding the Bregman Projection in Mirror Descent

Context: In mirror descent, the Bregman projection P_C^h plays a crucial role in steering the iterates to remain within the feasible set C, thereby adhering to the problem's constraints.

Bregman Projection Definition
The Bregman projection is defined by the process of minimizing the Bregman divergence with respect to a set C:

$$\begin{aligned} P_C^h(y) &= \arg\min_{x \in C} D_h(x, y) \\ &= \arg\min_{x \in C} \left\{ h(x) - (\nabla h(x^{(k)}) - \alpha_k g^{(k)})^T x \right\}, \end{aligned}$$

where $D_h(x, y)$ is the Bregman divergence, h is a strongly convex function, $g^{(k)}$ is the gradient at iteration k, and α_k is the learning rate.

Optimality Conditions and Iterative Steps in Mirror Descent

Overview: The mirror descent algorithm progresses through its iterations by adhering to specific optimality conditions. These conditions ensure that each step is efficiently guided towards minimizing the objective function.

Optimality Conditions

The key condition for each iteration k in mirror descent is represented by the equation:

$$\nabla g^{(k)} + \frac{1}{\alpha_k} \nabla h(y) - \frac{1}{\alpha_k} \nabla h(x^{(k)}) = 0,$$

where $\nabla g^{(k)}$ is the gradient of the loss function at the current iterate, α_k is the learning rate, and ∇h denotes the gradient of the mirror map function.

The Mirror Descent Algorithm: Detailed Explanation

Algorithm 1 Mirror Descent Algorithm

1: **Initialize:** Set initial point $x_1 \in \mathcal{X} \cap D$, step size η.
2: **for** $i \leftarrow 1, 2, \ldots$ **do**
3: $\quad i \leftarrow i + 1$.
4: $\quad$ Compute the incurred cost $f_i(x_i)$ and subgradient $g_i \in \partial f_i(x_i)$
5: $\quad$ Map from primal to dual space: Calculate $\hat{x}_i = \nabla\Phi(x_i)$
6: $\quad$ Perform gradient descent step in dual space: Update $\hat{y}_{i+1} = \hat{x}_i - \eta g_i$, adjusting for the learning rate η.
7: $\quad$ Inverse mapping to primal space: Find $y_{i+1} = \nabla\Phi^*(\hat{y}_{i+1})$, where Φ^* is the convex conjugate of Φ.
8: $\quad$ Project back to feasible region: Determine $x_{i+1} = \Pi^{\Phi}_{\mathcal{X}\cap D}(y_{i+1})$.
9: **end for**=0

References

Stephen Boyd and Lieven Vandenberghe.
Convex Optimization.
Cambridge University Press, 2004.

Alexander Shapiro, Andrzej Piotr Ruszczyński, and Darinka Dentcheva
Lectures on Stochastic Programming: Modeling and Theory.
SIAM Publication Library, 2021.