Accelerated Proximal Gradient Methods

Paul Tseng (2008)

December 6, 2024

Presented by William Francis

outline

Problem Statement

$$\min_{x \in \mathcal{E}} f(x) + P(x)$$

Assumptions

- $\mathcal{E}$ is a linear space and $\operatorname{dom} P \neq \emptyset$
- f is continuously differentiable
- ∇f is L-Lipschitz
- P is proximable

Terminology

- $f^P(x) = f(x) + P(x)$

Terminology

- $f^P(x) = f(x) + P(x)$
- $\ell_f(x; y) = f(y) + \langle \nabla f(y), x - y \rangle + P(x)$

Terminology

- $f^P(x) = f(x) + P(x)$
- $\ell_f(x; y) = f(y) + \langle \nabla f(y), x - y \rangle + P(x)$
- $D(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle$

Terminology

- $f^P(x) = f(x) + P(x)$
- $\ell_f(x; y) = f(y) + \langle \nabla f(y), x - y \rangle + P(x)$
- $D(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle$
- Example: $h(x) = \frac{1}{2}\|x\|^2 \rightarrow D(x, y) = \frac{1}{2}\|x - y\|^2$

Basic Proximal Gradient

$$x_{k+1} = (I + \alpha \partial P)^{-1}(I - \nabla f)x_k$$

- $\mathcal{O}(1/k)$ convergence
- $\mathcal{O}(n)$ memory

Provable Results

- Best acceleration: $\mathcal{O}(1/k^2)$ function value convergence

Provable Results

- Best acceleration: $\mathcal{O}(1/k^2)$ function value convergence
- Duality gap shrinks with bounds depending on choice of momentum term, θ_k (q^P is dual function).

$$0 \leq f^P(x_{k+1} - q^P(\bar{v}_k) \leq \theta_k^2 L \max_{x \in \mathrm{dom} P} D(x, z_0)$$

Overview

- Some algorithms may be better for particular applications

Overview

- Some algorithms may be better for particular applications
- Use “momentum” to accelerate

Overview

- Some algorithms may be better for particular applications
- Use “momentum” to accelerate
- Momentum decreases over time to hone in

Overview

- Some algorithms may be better for particular applications
- Use “momentum” to accelerate
- Momentum decreases over time to hone in
- Sometimes overshoots, creating small oscillations

Algorithm 1

$$y_k = (1 - \theta_k)x_k + \theta_k z_k$$
$$z_{k+1} = \arg\min_{x \in X_k} \{\ell_f(x; y_k) + \theta_k L D(x, z_k)\}$$
$$\hat{x}_{k+1} = (1 - \theta_k)x_k + \theta_k z_{k+1}$$

with the constraints

$$\frac{1 - \theta_{k+1}}{\theta_{k+1}^2} \leq \frac{1}{\theta_k^2}$$

$$\ell_f(x_{k+1}; y_k) + \frac{L}{2}\|x_{k+1} - y_k\|^2 \leq \ell_f(\hat{x}_{k+1}; y_k) + \frac{L}{2}\|\hat{x}_{k+1} - y_k \; Vert^2$$

Algorithm 1

$$\theta_k = (1/2)(\sqrt{\theta_{k-1}^4 + 4\theta_{k-1}^2} - \theta_{k-1}^2)$$

$$y_k = (1 - \theta_k)x_k + \theta_k z_k$$

$$z_{k+1} = (I + \frac{1}{\theta_k L}\partial P)^{-1}(z_k - \frac{1}{\theta_k L}\nabla f(y_k))$$

$$x_{k+1} = (1 - \theta_k)x_k + \theta_k z_{k+1}$$

- $\mathcal{O}(1/k^2)$ convergence
- $\mathcal{O}(n)$ memory

Algorithm 1

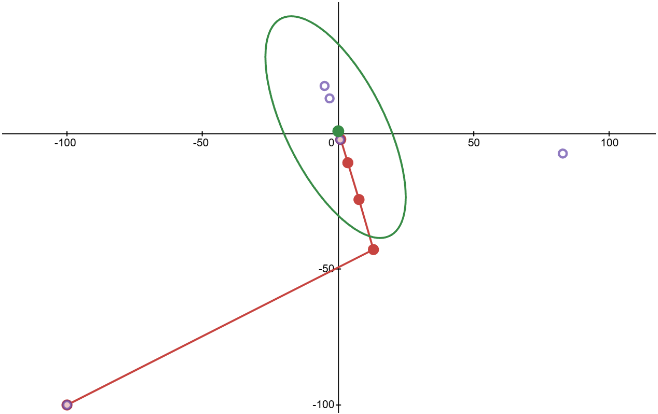


Figure: Red: x_k, Purple: z_k

Algorithm 2

$$y_k = x_k + \theta_k(\theta_{k-1}^{-1} - 1)(x_k - x_{k-1})$$
$$x_{k+1} = (I + \frac{1}{L}\partial P)^{-1}(I - \frac{1}{L}\nabla f)y_k$$

- This is the one we did in class, with $t_k = 1/\theta_k$.

Algorithm 3

$$y_k = (1 - \theta_k)x_k + \theta_k z_k$$

$$z_{k+1} = \arg\min_x \{\sum_{i=0}^{k} \frac{\ell_f(x; y_i}{\vartheta_i} + Lh(x)\}$$

$$x_{k+1} = (1 - \theta_k)x_k + \theta_k z_{k+1}$$

- Weighted sum of gradients from previous iterations
- Typically $\vartheta_k = \theta_k$, but technically can be relaxed.
- $\{\theta_k\}$ decreasing, so more recent terms weighted higher

Algorithm 3

$$y_k = (1-\theta_k)x_k + \theta_k z_k$$

$$z_{k+1} = (I + \frac{1}{L}(\sum_{i=0}^{k} \frac{1}{\vartheta_k})\partial P)^{-1}(-\frac{1}{L}\sum_{i=1}^{k} \nabla f(y_i))$$

$$x_{k+1} = (1-\theta_k)x_k + \theta_k z_{k+1}$$

- Weighted sum of gradients from previous iterations
- Typically $\vartheta_k = \theta_k$, but technically can be relaxed.
- $\{\theta_k\}$ decreasing, so more recent terms weighted higher

Algorithm 3

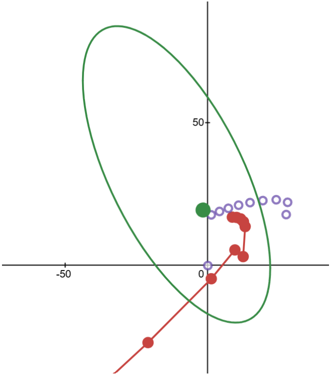


Figure: Red: x_k, Purple: z_k

Search Region Reduction

- $\forall w, x \in \mathrm{dom} P$ such that $f(x) \leq \inf f + \epsilon$, $f^P(w) + \epsilon \geq f^P(x) \geq \ell_f(x; w)$ by convexity.

Search Region Reduction

- $\forall w, x \in \text{dom}P$ such that $f(x) \leq \inf f + \epsilon$, $f^P(w) + \epsilon \geq f^P(x) \geq \ell_f(x; w)$ by convexity.
- So the half-space $\ell_f(w; x) - f^P(x) \leq \epsilon$ contains all ϵ-minimum points of f^P for any w.

Search Region Reduction

- $\forall w, x \in \mathrm{dom} P$ such that $f(x) \leq \inf f + \epsilon$, $f^P(w) + \epsilon \geq f^P(x) \geq \ell_f(x; w)$ by convexity.
- So the half-space $\ell_f(w; x) - f^P(x) \leq \epsilon$ contains all ϵ-minimum points of f^P for any w.
- So convex combinations of these half spaces do too

$$X_k = \{x : \sum_{i \in I_{k,j}} \alpha_{k,i}(\ell_f(x, w_{k,i}) - f^P(w_{k,i})) \leq \epsilon, j = 1, \ldots, n_k\}$$

Search Region Reduction

- $\forall w, x \in \text{dom}P$ such that $f(x) \leq \inf f + \epsilon$, $f^P(w) + \epsilon \geq f^P(x) \geq \ell_f(x; w)$ by convexity.
- So the half-space $\ell_f(w; x) - f^P(x) \leq \epsilon$ contains all ϵ-minimum points of f^P for any w.
- So convex combinations of these half spaces do too

$$X_k = \{x : \sum_{i \in I_{k,j}} \alpha_{k,i}(\ell_f(x, w_{k,i}) - f^P(w_{k,i})) \leq \epsilon, j = 1, ..., n_k\}$$

- Convex combinations of half-spaces are half-spaces

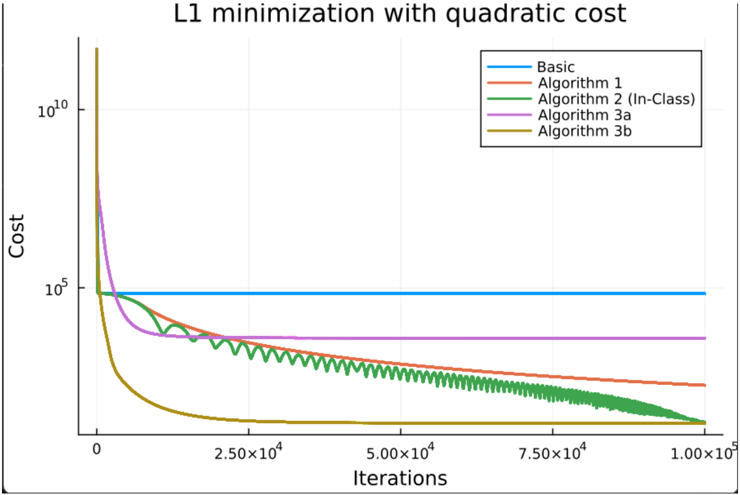
Search Region Reduction

- $\forall w, x \in \text{dom}P$ such that $f(x) \leq \inf f + \epsilon$, $f^P(w) + \epsilon \geq f^P(x) \geq \ell_f(x; w)$ by convexity.
- So the half-space $\ell_f(w; x) - f^P(x) \leq \epsilon$ contains all ϵ-minimum points of f^P for any w.
- So convex combinations of these half spaces do too

$$X_k = \{x : \sum_{i \in I_{k,j}} \alpha_{k,i}(\ell_f(x, w_{k,i}) - f^P(w_{k,i})) \leq \epsilon, j = 1, ..., n_k\}$$

- Convex combinations of half-spaces are half-spaces
- Half spaces relatively easy to search in, but still increases cost per iteration.

Not Strongly Convex



Strongly Convex