Bregman proximal methods for convex optimization

Javier Peña, Carnegie Mellon University
(joint work with D. Gutman)

IEEE-CCAC
Medellín, October 2019

Some motivation

Convex optimization

Problem of the form

$$\min_{x \in Q} f(x)$$

where $Q \subseteq \mathbb{R}^n$ and $f : Q \to \mathbb{R}$ are convex.

Many applications

- *Classic:*
  - linear programming models for production, logistics, etc.
  - quadratic programming models for portfolio construction
  - integer programming and combinatorial optimization
- *Modern:*
  - data science: support vector machines, regression, matrix completion
  - imaging science: compressive sensing
  - computational game theory

Recall: $Q \subseteq \mathbb{R}^n$ and $f : Q \to \mathbb{R}$ are convex if for all $x, y \in Q$ and $\lambda \in [0, 1]$

$$\lambda x + (1 - \lambda)y \in Q \text{ and } f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$$

Incomplete & biased history of algorithms for convex opt

- Late 20th century (1980s–2000)
  - interior-point (second-order) methods
  - strong theory, successful implementations, high accuracy
  - semidefinite & second-order programming
- Early 21st century (2000–now)
  - large-scale problems
  - modest accuracy is often acceptable
  - resurgence of first-order methods: **the topic of this talk**

Preamble: some iconic algorithms

Unconstrained convex minimization

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable and consider the problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

Gradient descent (GD)

$$\begin{aligned} &\text{pick } \ t_k > 0 \\ &x_{k+1} = x_k - t_k \nabla f(x_k) \end{aligned}$$

Accelerated gradient (AG)

$$\begin{aligned} &\text{pick } \ \beta_k \geq 0, \ t_k > 0 \\ &y_k = x_k + \beta_k (x_k - x_{k-1}) \\ &x_{k+1} = y_k - t_k \nabla f(y_k) \end{aligned}$$

Composite convex minimization

Consider the problem

$$\min_{x\in\mathbb{R}^n} \{f(x) + \psi(x)\}$$

where $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is differentiable and convex, and $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is closed and convex with $\mathrm{dom}(\psi) \subseteq \mathrm{dom}(f)$.

Let Prox_t be the following *proximal map*

$$\mathsf{Prox}_t(x) := \operatorname*{argmin}_{z\in\mathbb{R}^n} \left\{ \frac{1}{2t}\|z - x\|^2 + \psi(z) \right\}.$$

Observe: if $\psi = \delta_Q$ then

$$\min_{x\in\mathbb{R}^n}\{f(x) + \psi(x)\} \Leftrightarrow \min_{x\in Q} f(x)$$

and for all $t > 0$

$$\mathsf{Prox}_t(x) = \mathsf{Proj}_Q(x) = \operatorname*{argmin}_{z\in Q} \|z - x\|.$$

Proximal gradient and accelerated proximal gradient

Consider the problem

$$\min_{x \in \mathbb{R}^n} \{f(x) + \psi(x)\}.$$

Proximal gradient (PG)

$$\begin{aligned}
&\text{pick } \ t_k > 0 \\
&x_{k+1} = \mathsf{Prox}_{t_k}(x_k - t_k \nabla f(x_k))
\end{aligned}$$

Accelerated proximal gradient (APG)

$$\begin{aligned}
&\text{pick } \ \beta_k \geq 0, \ t_k > 0 \\
&y_k = x_k + \beta_k(x_k - x_{k-1}) \\
&x_{k+1} = \mathsf{Prox}_{t_k}(y_k - t_k \nabla f(y_k))
\end{aligned}$$

Choice of stepsize

Consider the generic update

$$z_+ = \mathsf{Prox}_t(y - t\nabla f(y)).$$

Observe

$$\begin{aligned}&\mathsf{Prox}_t(y - t\nabla f(y))\\ &\qquad = \operatorname*{argmin}_{z\in\mathbb{R}^n} \left\{ f(y) + \langle \nabla f(y), z-y\rangle + \frac{1}{2t}\|z-y\|^2 + \psi(z)\right\}.\end{aligned}$$

It makes sense to choose t so that $z_+ = \mathsf{Prox}_t(y - t\nabla f(y))$ satisfies

$$f(z_+) + \psi(z_+) \le f(y) + \langle \nabla f(y), z_+ - y\rangle + \frac{1}{2t}\|z_+ - y\|^2 + \psi(z_+)$$

or equivalently

$$f(z_+) - f(y) - \langle \nabla f(y), z_+ - y\rangle \le \frac{1}{2t}\|z_+ - y\|^2.$$

Bregman distance and L-smoothness

The latter condition can be restated as

$$D_f(z_+, y) \le \frac{1}{2t}\|z_+ - y\|^2$$

where D_f is the following *Bregman distance* generated by f

$$D_f(z, y) := f(z) - f(y) - \langle \nabla f(y), z - y\rangle.$$

L-smoothness

We say that f is L-smooth if for all $z, y \in \mathrm{dom}(f)$

$$D_f(z, y) \le \frac{L}{2}\|z - y\|^2.$$

In this case the condition at the top holds for $t = 1/L$.

Fact: f is L-smooth if ∇f is L-Lipschitz.

Convergence of PG

PG: solve $\min_x \{f(x) + \psi(x)\}$ via

$$x_{k+1} = \mathsf{Prox}_{t_k}(x_k - t_k \nabla f(x_k)).$$

Theorem

If the stepsizes t_k satisfy

$$D_f(x_{k+1}, x_k) \leq \frac{1}{2t_k} \|x_{k+1} - x_k\|^2$$

then for all $\bar{x} \in \mathrm{argmin}_x \{f(x) + \psi(x)\}$ the PG iterates satisfy

$$f(x_k) + \psi(x_k) - (f(\bar{x}) + \psi(\bar{x})) \leq \frac{\|x_0 - \bar{x}\|^2}{2 \sum_{i=0}^{k-1} t_i}.$$

In particular, if each $t_k \geq 1/L > 0$ then

$$f(x_k) + \psi(x_k) - (f(\bar{x}) + \psi(\bar{x})) \leq \frac{L \cdot \|x_0 - \bar{x}\|^2}{2k}.$$

Convergence of APG

APG: solve $\min_x \{f(x) + \psi(x)\}$ via

$$y_k = x_k + \beta_k(x_k - x_{k-1})$$
$$x_{k+1} = \mathsf{Prox}_{t_k}(y_k - t_k \nabla f(y_k))$$

Theorem (Beck & Teboulle 2009, Nesterov 2013)

Suppose $\beta_k = \frac{k-1}{k+2}$ *and the stepsizes* t_k *satisfy* $t_k \geq 1/L > 0$ *and*

$$D_f(x_{k+1}, y_k) \leq \frac{1}{2t_k}\|x_{k+1} - y_k\|^2.$$

Then for all $\bar{x} \in \mathsf{argmin}_x\{f(x) + \psi(x)\}$ *the APG iterates satisfy*

$$f(x_k) + \psi(x_k) - (f(\bar{x}) + \psi(\bar{x})) \leq \frac{2L \cdot \|x_0 - \bar{x}\|^2}{(k+1)^2}.$$

Main story: Bregman proximal methods

Proximal map again

Observe

$$
\begin{aligned}
&\mathsf{Prox}_t(x - t\nabla f(x)) \\
&\qquad = \operatorname*{argmin}_{y\in\mathbb{R}^n} \left\{ f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2t}\|y-x\|^2 + \psi(y) \right\} \\
&\qquad = \operatorname*{argmin}_{y\in\mathbb{R}^n} \left\{ \langle \nabla f(x), y\rangle + \psi(y) + \frac{1}{2t}\|y-x\|^2 \right\}.
\end{aligned}
$$

Also get $\mathcal{O}(1/k)$ and $\mathcal{O}(1/k^2)$ convergence of proximal gradient methods when f is L-smooth:

$$
D_f(y, x) \le \frac{L}{2}\|y - x\|^2.
$$

The above can be relaxed and extended.

Bregman proximal map

Let $h : \mathbb{R}^n \to \mathbb{R}$ be a differentiable convex *reference* function.

The *Bregman distance* associated to h is

$$D_h(y, x) := h(y) - h(x) - \langle \nabla h(x), y - x \rangle.$$

This distance defines the following *Bregman proximal map*

$$g \mapsto \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ \langle g, y \rangle + \psi(y) + \frac{1}{t} D_h(y, x) \right\}$$

The previous *Euclidean* proximal map corresponds to the squared Euclidean norm reference function

$$h(x) = \frac{\|x\|^2}{2} \rightsquigarrow D_h(y, x) = \frac{\|y - x\|^2}{2}.$$

Most of what we discussed for Euclidean proximal methods extends to Bregman proximal methods.

Bregman proximal gradient

Consider the problem

$$\min_{x \in \mathbb{R}^n} \{f(x) + \psi(x)\},$$

and suppose $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a reference function.

Bregman proximal gradient (BPG)

$$\text{pick } \; t_k > 0$$

$$x_{k+1} = \operatorname*{argmin}_{z \in \mathbb{R}^n} \left\{ \langle \nabla f(x_k), z \rangle + \psi(z) + \frac{1}{t_k} D_h(z, x_k) \right\}$$

Why use Bregman proximal methods?

- The Bregman proximal template provides a lot more flexibility.
- The additional freedom to choose h can facilitate the computation of the proximal mapping. For instance for $x \in \Delta_{n-1} := \{x \in \mathbb{R}^n_+ : \|x\|_1 = 1\}$ the map

  $$g \mapsto \operatorname*{argmin}_{y \in \Delta_{n-1}} \{\langle g, y \rangle + D_h(y, x)\}$$

  is more easily computable for $h(x) = \sum_{i=1}^n x_i \log(x_i)$.
- The usual L-smoothness assumption for convergence can be replaced by a *relative* L-smoothness that holds more broadly.

Example: D-optimal design problem (min-vol enclosing ellipsoid)

$$\min_{x \in \Delta_{n-1}} \; -\log(\det(HXH^\mathsf{T}))$$

where $X = \text{Diag}(x)$ and $H \in \mathbb{R}^{m \times n}$ with $m < n$.

Example: Poisson linear inverse problem

$$\min_{x \in \mathbb{R}^n_+} \; D_{KL}(b, Ax)$$

where $b \in \mathbb{R}^n_{++}$ and $A \in \mathbb{R}^{m \times n}_+$ with $m > n$ and $D_{KL}(\cdot, \cdot)$ is the Kullback-Leibler divergence, that is, the Bregman distance associated to $x \mapsto \sum_{i=1}^n x_i \log(x_i)$.

We could tackle the above two problems via Euclidean proximal methods. However, they are more amenable to Bregman proximal methods with the *Burg entropy* reference function

$$h(x) = -\sum_{i=1}^{n} \log(x_i).$$

Accelerated Bregman proximal gradient (ABPG)

(Gutman-P 2018)

Generate sequences x_k, y_k, z_k for $k = 0, 1, \ldots$ as follows:

$$
\begin{aligned}
&\text{pick } t_k > 0 \\
&z_{k+1} = \operatorname*{argmin}_{z \in \mathbb{R}^n} \left\{ \langle \nabla f(y_k), z \rangle + \psi(z) + \frac{1}{t_k} D_h(z, z_k) \right\} \\
&x_{k+1} = \frac{\sum_{i=0}^{k} t_i z_{i+1}}{\sum_{i=0}^{k} t_i} \\
&y_{k+1} = \frac{\sum_{i=0}^{k} t_i z_{i+1} + t_{k+1} z_{k+1}}{\sum_{i=0}^{k+1} t_i}
\end{aligned}
$$

Related work by Hanzely-Richtarik-Xiao (2018).

Fenchel duality

Convex conjugate

For $\phi : \mathbb{R}^n \to \mathbb{R}$ let $\phi^* : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be defined via

$$\phi^*(u) = \sup_{x \in \mathbb{R}^n} \{\langle u, x\rangle - \phi(x)\}.$$

Observe: ϕ^* is always convex, even when ϕ is not.

Consider the *primal* problem

$$\min_x \{f(x) + \psi(x)\}.$$

The corresponding *Fenchel dual* problem is

$$\max_u \{-f^*(u) - \psi^*(-u)\}.$$

Fenchel duality

Recall: $\phi^*(u) = \sup_x\{\langle u, x\rangle - \phi(x)\}$.

Weak duality

For all $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^n$

$$f(x) + \psi(x) \geq -f^*(u) - \psi^*(-u).$$

Proof.

$f(x) + f^*(u) + \psi(x) + \psi^*(-u) \geq \langle u, x\rangle + \langle -u, x\rangle = 0.$ $\square$

Fenchel duality

Recall: $\phi^*(u) = \sup_x\{\langle u, x\rangle - \phi(x)\}$.

An approach to show convergence

Suppose an algorithm generates sequences x_k, v_k, w_k such that

$$f(x_k) + \psi(x_k) \le -f^*(v_k) - \psi^*(w_k) - d_k^*(-v_k - w_k)$$

for some sequence of "distance" functions $d_k : \mathbb{R}^n \to \mathbb{R}$.

Then for all $\bar{x} \in \operatorname{argmin}_x\{f(x) + \psi(x)\}$ we have

$$\begin{aligned} & f(x_k) + \psi(x_k) \\ & \quad \le -\langle v_k, \bar{x}\rangle + f(\bar{x}) - \langle w_k, \bar{x}\rangle + \psi(\bar{x}) + \langle v_k + w_k, \bar{x}\rangle + d_k(\bar{x}) \\ & \quad = f(\bar{x}) + \psi(\bar{x}) + d_k(\bar{x}). \end{aligned}$$

That is, $f(x_k) + \psi(x_k) - (f(\bar{x}) + \psi(\bar{x})) \le d_k(\bar{x})$.

A key lemma

Suppose $y_k, z_k \in \mathrm{ri}(\mathrm{dom}(h)) \cap \mathrm{dom}(\psi), g_k := \nabla f(y_k)$, and $t_k > 0$ satisfy

$$z_{k+1} = \operatorname*{argmin}_{z \in \mathbb{R}^n} \left\{ \langle g_k, z \rangle + \psi(z) + \frac{1}{t_k} D_h(z, z_k) \right\}$$

for $k = 0, 1, 2, \ldots$.

Via the optimality conditions rewrite above as

$$g_k + g_k^{\psi} + \frac{1}{t_k}(\nabla h(z_{k+1}) - \nabla h(z_k)) = 0$$

for some $g_k^{\psi} \in \partial \psi(z_{k+1})$.

A key lemma

Let

$$v_k := \frac{\sum_{i=0}^k t_i g_i}{\sum_{i=0}^k t_i}, \; w_k := \frac{\sum_{i=0}^k t_i g_i^\psi}{\sum_{i=0}^k t_i}.$$

Lemma

Suppose $y_k, z_k, g_k, g_k^\psi, t_k$ *and* v_k, w_k *are as above. Then*

$$\begin{aligned}
&\frac{\sum_{i=0}^k t_i(f(z_{i+1}) + \psi(z_{i+1}) - D_f(z_{i+1}, y_i)) + D_h(z_{i+1}, z_i)}{\sum_{i=0}^k t_i} \\
&= -\frac{\sum_{i=0}^k t_i(f^*(g_i) + \psi^*(g_i^\psi))}{\sum_{i=0}^k t_i} - d_k^*(-v_k - w_k) \\
&\le -f^*(v_k) - \psi^*(w_k) - d_k^*(-v_k - w_k)
\end{aligned}$$

where

$$d_k(z) := \frac{1}{\sum_{i=0}^k t_i} D_h(z, z_0).$$

Bregman proximal gradient (BPG)

In this case

$$x_{k+1} = \operatorname*{argmin}_{z\in\mathbb{R}^n} \left\{ \langle \nabla f(x_k), z\rangle + \psi(z) + \frac{1}{t_k} D_h(z, x_k) \right\}$$

Theorem (Gutman-P 2018)

Suppose each t_i is such that

$$D_f(x_{i+1}, x_i) \le \frac{1}{t_i} D_h(x_{i+1}, x_i). \tag{DC}$$

Then for $\bar{x} \in \operatorname*{argmin}_{x\in\mathbb{R}^n}\{f(x) + \psi(x)\}$ the BPG iterates satisfy

$$f(x_{k+1}) + \psi(x_{k+1}) - (f(\bar{x}) + \psi(\bar{x})) \le \frac{1}{\sum_{i=0}^k t_i} D_h(\bar{x}, x_0).$$

Proof of Theorem.

In this case we can apply Lemma to $x_k = y_k = z_k$ and get

$$\frac{\sum_{i=0}^{k} t_i(f(x_{i+1}) + \psi(x_{i+1}) - D_f(x_{i+1}, x_i)) + D_h(x_{i+1}, x_i)}{\sum_{i=0}^{k} t_i}$$
$$\leq -f^*(v_k) - \psi^*(w_k) - d_k^*(-v_k - w_k)$$

Next, (DC) implies $f(x_{i+1}) + \psi(x_{i+1}) \leq f(x_i) + \psi(x_i)$ and

$$f(x_{k+1}) + \psi(x_{k+1}) \leq -f^*(v_k) - \psi^*(w_k) - d_k^*(-v_k - w_k).$$

To finish, observe that for all $\bar{x} \in \text{argmin}_{x\in\mathbb{R}^n}\{f(x) + \psi(x)\}$

$$\begin{aligned}
&-f^*(v_k) - \psi^*(w_k) - d_k^*(-v_k - w_k) \\
&\quad \leq -\langle v_k, \bar{x}\rangle + f(\bar{x}) - \langle w_k, \bar{x}\rangle + \psi(\bar{x}) + \langle v_k + w_k, \bar{x}\rangle + d_k(\bar{x}) \\
&\quad = f(\bar{x}) + \psi(\bar{x}) + d_k(\bar{x}) \\
&\quad = f(\bar{x}) + \psi(\bar{x}) + \frac{1}{\sum_{i=0}^{k} t_i} D_h(\bar{x}, x_0).
\end{aligned}$$

Relative smoothness

Suppose f, h are convex and differentiable on Q. We say that f is L-smooth relative to h on Q if for all $x, y \in Q$

$$D_f(y, x) \leq L D_h(y, x).$$

(Nguyen 2012, Bauschke et al. 2017, Lu et al. 2018)

If f is L-smooth relative to h on $\operatorname{dom}(\psi)$ then (DC) holds for $t_i = 1/L,\ i = 0, 1, \ldots, k-1$ and BPG iterates satisfy

$$f(x_k) + \psi(x_k) - (f(\bar{x}) + \psi(\bar{x})) \leq \frac{L D_h(\bar{x}, x_0)}{k}.$$

Recover results by Bauschke-Bolte-Teboulle (2017) and by Lu-Freund-Nesterov (2018).

This extends the $\mathcal{O}(1/k)$ convergence rate of PG.

Accelerated Bregman proximal gradient (ABPG)

Generate sequences x_k, y_k, z_k as follows:

$$z_{k+1} = \operatorname*{argmin}_{z \in \mathbb{R}^n} \left\{ \langle \nabla f(y_k), z \rangle + \psi(z) + \frac{1}{t_k} D_h(z, z_k) \right\}$$

$$x_{k+1} = \frac{\sum_{i=0}^{k} t_i z_{i+1}}{\sum_{i=0}^{k} t_i}$$

$$y_{k+1} = \frac{\sum_{i=0}^{k} t_i z_{i+1} + t_{k+1} z_{k+1}}{\sum_{i=0}^{k+1} t_i}.$$

By letting $\theta_k := \frac{t_k}{\sum_{i=0}^{k} t_i}$ the last two equations can be rewritten as

$$\begin{aligned} x_{k+1} &= (1-\theta_k) x_k + \theta_k z_{k+1} \\ y_{k+1} &= (1-\theta_{k+1}) x_{k+1} + \theta_{k+1} z_{k+1} \\ &= x_{k+1} + \frac{\theta_{k+1}(1-\theta_k)}{\theta_k}(x_{k+1} - x_k) \end{aligned}$$

Accelerated Bregman proximal gradient (ABPG)

Theorem (Gutman-P 2018)

Suppose each t_i and θ_i are such that

$$D_f(x_{i+1}, y_i) - (1-\theta_i) D_f(x_i, y_i) \le \frac{\theta_i}{t_i} D_h(z_{i+1}, z_i). \qquad \text{(ADC)}$$

Then for $\bar{x} \in \bar{X} := \operatorname*{argmin}_{x \in \mathbb{R}^n} \{f(x) + \psi(x)\}$ the ABPG iterates satisfy

$$f(x_{k+1}) + \psi(x_{k+1}) - (f(\bar{x}) + \psi(\bar{x})) \le \frac{1}{\sum_{i=0}^{k} t_i} D_h(\bar{x}, x_0).$$

Proof.

Similar to previous one for BPG: use lemma & Fenchel duality. □

Compare (ADC) condition for ABPG with (DC) condition for BPG:

$$D_f(x_{i+1}, x_i) \le \frac{1}{t_i} D_h(x_{i+1}, x_i). \qquad \text{(DC)}$$

Relative smoothness revisited

How much can we accelerate?

Choose $t_k > 0$ or equivalently $\theta_k = \frac{t_k}{\sum_{i=0}^k t_i}$ as large as possible so that (ADC) holds. How large can we choose it?

(L, γ) relative smoothness

Say that f is (L, γ)-smooth relative to h on Q if for all $x, y, z, \tilde{z} \in Q$ and $\theta \in [0, 1]$

$$D_f((1-\theta)x + \theta\tilde{z}, (1-\theta)x + \theta z) \leq L\theta^\gamma D_h(\tilde{z}, z).$$

Observe

In Euclidean case L-relative smoothness yields $(L, 2)$-relative smoothness. In general this does not hold but "almost"...

Accelerated Bregman proximal gradient

Theorem (Gutman-P 2018)

Suppose f is (L,γ)-smooth relative to h on ri$(\mathrm{dom}(h)) \cap \mathrm{dom}(\psi)$ for some $L > 0$ and $\gamma > 0$.

Then the stepsizes t_k can be chosen judiciously so that the ABPG iterates satisfy

$$f(x_{k+1}) + \psi(x_{k+1}) - (f(\bar{x}) + \psi(\bar{x})) \leq \left(\frac{\gamma}{k+\gamma}\right)^{\gamma} L D_h(\bar{X}, x_0).$$

Recover iconic $\mathcal{O}(1/k^2)$ rate when $h(x) = \frac{1}{2}\|x\|^2$ and f is L-smooth.

Accelerated Bregman proximal gradient

For implementation purposes: pick $\theta_k = \frac{t_k}{\sum_{i=0}^k t_i}$ as large as possible so that (ADC) holds. Pick θ_k of the form

$$\theta_k = \frac{\gamma_k}{k + \gamma_k}$$

via backtracking on γ_k. If all $\gamma_k \geq \gamma > 0$ then we get

$$f(x_{k+1}) + \psi(x_{k+1}) - (f(\bar{x}) + \psi(\bar{x})) \leq \left(\frac{\gamma}{k+\gamma}\right)^{\gamma} L D_h(\bar{X}, x_0).$$

If we can do the above with $\gamma = 2$ we recover $\mathcal{O}(1/k^2)$ rate. This happens when $h(x) = \frac{1}{2}\|x\|^2$.

Numerical experiments

BPG-LS and ABPG-LS implementations

- Line-search to choose t_k in BPG so that (DC) holds.
- Likewise for t_0 and $\theta_k \in (0,1)$ in ABPG to ensure (ADC).
- Pick $\theta_k \in (0,1)$ of the form $\theta_k = \frac{\gamma_k}{k+\gamma_k}$.

ABPG: use educated guess for t_0 and $\theta_k = 2/(k+2)$.

Problem instances

- D-optimal design: $\min_{x\in\Delta_{n-1}} \; -\log(\det(HXH^\mathsf{T}))$
- Poisson linear inverse: $\min_{x\in\mathbb{R}^n_+} \; D_{KL}(b, Ax)$

In both cases use reference function

$$h(x) = -\sum_{i=1}^{n} \log(x_i).$$

Bregman proximal mappings are easily computable in both cases.

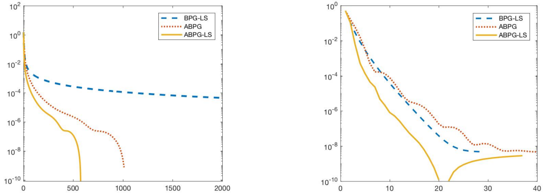

Figure: Suboptimality gap for 100×250 and 200×300 random instances of D-optimal design.

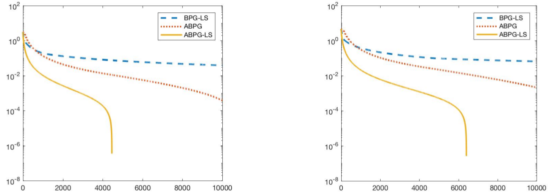

Figure: Suboptimality gap for 250×100 and 300×200 random instances of the Poisson linear inverse problem.

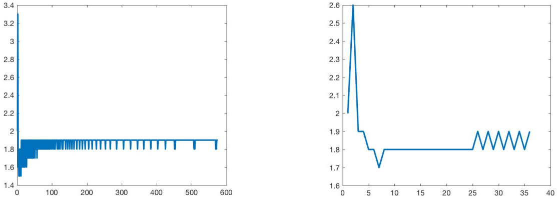

Figure: Sequence $\{\gamma_k : k = 1, 2, \dots\}$ in ABPG-LS for typical instances of D-design optimal problem.

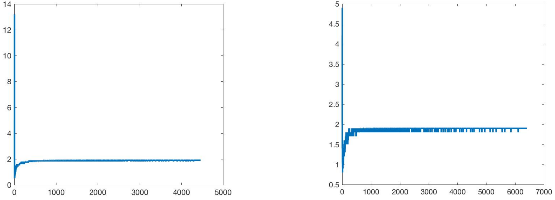

Figure: Sequence $\{\gamma_k : k = 1, 2, \dots\}$ in ABPG-LS for typical instances of Poisson linear inverse problem.

Conclusions

- Analysis of Bregman proximal methods via Fenchel duality.
  Key observation: algorithms generate x_k, v_k, w_k such that

$$f(x_{k+1}) + \psi(x_{k+1}) \leq -f^*(v_k) - \psi^*(w_k) - d_k^*(-v_k - w_k).$$

- Other related developments that we did not discuss:
  - Proximal subgradient method when f is non-differentiable
  - Linear convergence via restarting
  - Analogous results for conditional gradient
- Current/future work
  - Saddle-point problems
  - Stochastic first-order methods
  - More computational experiments
  - Role of γ in accelerated Bregman proximal methods

Main references

- Bauschke, Bolte, Teboulle (2017), "A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications"
- Lu, Freund, Nesterov (2018), "Relatively smooth convex optimization by first-order methods, and applications"
- Gutman and Peña (2018), "Convergence rates of proximal gradient methods via the convex conjugate"
- Hanzely, Richtarik, Xiao (2018), "Accelerated Bregman proximal gradient methods for relatively smooth convex optimization"
- Gutman and Peña (2018), "A unified framework for Bregman proximal methods: subgradient, gradient, and accelerated gradient schemes"
- Teboulle (2018), "A simplified view of first-order methods for optimization"