Perturbed Fenchel Duality and First-Order Methods

Javier Peña, Carnegie Mellon University
joint work with D. Gutman, Texas Tech

Rutgers University, March 2022
Preamble: some motivation
Convex optimization

Problem of the form

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

where $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ and $C \subseteq \text{dom}(f)$ 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: equilibria computation
Incomplete & biased history

- Late 20th century (1980s–2000)
  - interior-point (second-order) methods
  - strong theory, successful code, high accuracy
  - semidefinite & second-order programming
  - elaborate algorithms and implementations for generic problems

- Early 21st century (2000–now)
  - large-scale problems
  - modest accuracy is often acceptable
  - resurgence of first-order methods – topic of this talk
  - simpler algorithms and implementations for specific problems
Popular formats

**Simple constraints**

\[
\min_{x \in C} f(x)
\]

where \( C \) is a “simple” set.

**Composite minimization**

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

where \( f, \psi \) are convex and \( \psi \) has some special structure.

Composite case subsumes the constrained case by taking \( \psi := \delta_C \)
where

\[
\delta_C(x) = \begin{cases} 
0 & \text{if } x \in C \\
\infty & \text{if } x \notin C.
\end{cases}
\]
Iconic algorithms for \( \min_{x \in C} f(x) \)

Let \( \Pi_C : \mathbb{R}^n \rightarrow C \) denote the orthogonal projection onto \( C \).

**Projected subgradient method (SG)**

Pick \( g_k \in \partial f(x_k) \) and \( t_k > 0 \)

\[
x_{k+1} = \Pi_C(x_k - t_k g_k)
\]

**Projected gradient descent (GD)**

Pick \( t_k > 0 \)

\[
x_{k+1} = \Pi_C(x_k - t_k \nabla f(x_k))
\]

**Conditional gradient (CG)**

\[
s_k = \arg\min_{s \in C} \langle \nabla f(x_k), s \rangle
\]

Pick \( \theta_k \in [0, 1] \)

\[
x_{k+1} = x_k + \theta_k (s_k - x_k)
\]
Iconic algorithms for \( \min_{x \in \mathbb{R}^n} \{ f(x) + \psi(x) \} \)

Suppose the following proximal mapping is computable for all \( t > 0 \)

\[
g \mapsto \text{Prox}_t(g) := \arg\min_{y \in \mathbb{R}^n} \left\{ \psi(y) + \frac{1}{2t} \| y - g \|^2 \right\}
\]

Observe: if \( \psi = \delta_C \) then \( \text{Prox}_t = \Pi_C \) for all \( t > 0 \).

**Proximal gradient (PG)**

pick \( t_k > 0 \)

\[
x_{k+1} = \text{Prox}_{t_k}(x_k - t_k \nabla f(x_k))
\]

**Fast proximal gradient (FPG)**

pick \( t_k > 0 \) and \( \beta_k \)

\[
y_k = x_k + \beta_k(x_k - x_{k-1})
\]

\[
x_{k+1} = \text{Prox}_{t_k}(y_k - t_k \nabla f(y_k))
\]

(Nesterov (1984), Beck-Teboulle (2009), Nesterov (2013),...)
Bregman proximal gradient for \( \min_{x \in \mathbb{R}^n} \{ f(x) + \psi(x) \} \)

Suppose \( h \) is a convex and differentiable reference function and the following proximal mapping is computable for all \( t > 0 \)

\[
(g, x) \mapsto \arg\min_{y \in \mathbb{R}^n} \left\{ \psi(y) + \langle g, y \rangle + \frac{1}{t} D_h(y, x) \right\}
\]

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

**Bregman proximal gradient (BPG)**

pick \( t_k > 0 \)

\[
x_{k+1} = \arg\min_{y \in \mathbb{R}^n} \left\{ \psi(y) + \langle \nabla f(x_k), y \rangle + \frac{1}{t_k} D_h(y, x_k) \right\}
\]

**Special case**

When \( h(x) = \|x\|^2_2 / 2 \), the Bregman proximal gradient becomes the previous (Euclidean) proximal gradient.
Convergence properties

Under suitable assumptions of smoothness and choice of stepsizes:

| Algorithm          | Convergence rate |
|--------------------|------------------|
| SG                 | $O(1/\sqrt{k})$  |
| GD, CG, PG, BPG    | $O(1/k)$         |
| FPG                | $O(1/k^2)$       |

Question

So many algorithms and so many convergence results. Could all of the above be “unified”?

*Answer*: YES, via *perturbed* Fenchel duality.
Theme

- A generic first-order meta-algorithm satisfies a perturbed Fenchel duality property.

- The first-order meta-algorithm includes as special cases: conditional gradient, proximal gradient, fast and universal proximal gradient, proximal subgradient.

- The perturbed Fenchel duality property yields concise derivations of the best-known convergence rates for each of these algorithms.
Perturbed Fenchel Duality
The Fenchel conjugate

Suppose \( f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\} \). The Fenchel conjugate of \( f \) is:

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

Fenchel-Young inequality

For all \( x, u \in \mathbb{R}^n \)

\[
f^*(u) + f(x) \geq \langle u, x \rangle,
\]

and the equality holds if and only if \( u \in \partial f(x) \).

Recall

\[
\partial f(x) = \{u \in \mathbb{R}^n : f(y) \geq f(x) + \langle u, y - x \rangle \text{ for all } y \in \mathbb{R}^n\}.
\]
Fenchel duality

The Fenchel dual of \( \min_{x \in \mathbb{R}^n} \{ f(x) + \psi(x) \} \) is

\[
\max_{u \in \mathbb{R}^n} \{-f^*(u) - \psi^*(-u)\}
\]

Weak duality

For all \( x, u \in \mathbb{R}^n \)

\[
f(x) + \psi(x) + f^*(u) + \psi^*(-u) \geq 0.
\]

Thus \( \bar{x}, \bar{u} \in \mathbb{R}^n \) are \( \epsilon \)-optimal if

\[
f(\bar{x}) + \psi(\bar{x}) + f^*(\bar{u}) + \psi^*(-\bar{u}) \leq \epsilon.
\]
Perturbed Fenchel duality

Gist of my story
First-order meta-algorithm generates $x_k, u_k \in \mathbb{R}^n$ such that

$$f(x_k) + \psi(x_k) + f^*(u_k) + (\psi + d_k)^*(-u_k) \leq \epsilon_k$$

for some $\epsilon_k \geq 0$ and $d_k : \mathbb{R}^n \to \mathbb{R}_+$ both converging to zero.

Observe
For all $x \in \mathbb{R}^n$ we have

$$f^*(u_k) + (\psi + d_k)^*(-u_k) \geq -f(x) - \psi(x) - d_k(x)$$

and thus perturbed Fenchel duality implies that

$$f(x_k) + \psi(x_k) - (f(x) + \psi(x)) \leq d_k(x) + \epsilon_k.$$
First-Order Meta-Algorithm
First-order meta-algorithm

Want to solve \( \min_x \{ f(x) + \psi(x) \} \).

Key ingredient

Let \( h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\} \) be a convex and differentiable reference function. Let \( D_h \) denote the Bregman distance

\[
D_h(y, x) = h(y) - h(x) - \langle \nabla h(x), y - x \rangle.
\]

Key assumption

The following proximal mapping is computable for all \( t > 0 \):

\[
(g, s_{-}) \mapsto \arg\min_s \left\{ \langle g, s \rangle + \psi(s) + \frac{1}{t} D_h(s, s_{-}) \right\}.
\]

Example

\( h(x) = \| x \|_2^2 / 2 \leadsto D_h(y, x) = \| y - x \|_2^2 / 2. \)
First-order meta-algorithm

Want to solve $\min_x \{ f(x) + \psi(x) \} \iff \min_x F(x)$ for $F := f + \psi$.

First-order meta-algorithm

- pick $s_{-1} \in \text{dom}(\psi)$
- for $k = 0, 1, \ldots$
  - pick $y_k \in \text{dom}(\partial f)$ and $g_k \in \partial f(y_k)$
  - pick $t_k > 0$
  - pick $s_k \in \text{argmin}_s \left\{ \langle g_k, s \rangle + \psi(s) + \frac{1}{t_k} D_h(s, s_{k-1}) \right\}$
- end for

Key component

Flexibly-selected sequence $y_k \in \text{dom}(f)$.

Specific choices of $y_k$: conditional gradient, Bregman proximal (sub)gradient, fast and universal Bregman proximal gradient.
Main Theorem

Let

$$x_k := \frac{\sum_{i=0}^{k-1} t_is_i}{\sum_{i=0}^{k-1} t_i}, u_k := \frac{\sum_{i=0}^{k-1} ti_g_i}{\sum_{i=0}^{k-1} t_i}, d_k(s) := \frac{D_h(s, s_{-1})}{\sum_{i=0}^{k-1} t_i}, \theta_k := \frac{t_k}{\sum_{i=0}^{k} t_i}.$$  

Theorem

The iterates generated by the above meta-algorithm satisfy

$$f(x_k) + \psi(x_k) + f^*(u_k) + (\psi + d_k)^*(-u_k) \leq \sum_{i=0}^{k-1} t_i(\mathcal{D}_F(x_i, s_i, \theta_i) + D_f(s_i, y_i)) - D_h(s_i, s_{i-1})$$

for (recall $F = f + \psi$)

$$\mathcal{D}_F(x, s, \theta) := \frac{F(x + \theta(s - x)) - (1 - \theta)F(x) - \theta F(s)}{\theta}.$$
Convergence of Iconic First-Order Algorithms
Conditional gradient

Want to solve \( \min_x \{ f(x) + \psi(x) \} \). Suppose \( f \) is differentiable and
\[
g \mapsto \partial \psi^*(-g) = \arg\min \{ \langle g, x \rangle + \psi(x) \}
\]
is computable.

Conditional gradient

- pick \( x_0 \in \text{dom}(f) \)
- for \( k = 0, 1, \ldots \)
  - pick \( s_k \in \arg\min_s \{ \langle \nabla f(x_k), s \rangle + \psi(s) \} \)
  - pick \( \theta_k \in [0, 1] \)
  - let \( x_{k+1} := (1 - \theta_k)x_k + \theta_k s_k \)
end for

This is the first-order meta-algorithm for
\[
s_{-1} = x_0, \ y_k = x_k, \ g_k = \nabla f(x_k), \ h \equiv 0,
\]
and \( t_k > 0 \) such that \( \theta_k = \frac{t_k}{\sum_{i=1}^k t_i} \).

(Mild assumption: \( \theta_0 = 1 \), and \( \theta_k \in (0, 1) \) for \( k \geq 1 \).)
Conditional gradient

Main Theorem yields

\[ f(x_k) + \psi(x_k) + f^*(u_k) + \psi^*(-u_k) \leq \frac{\sum_{i=0}^{k-1} t_i D(x_i, s_i, \theta_i)}{\sum_{i=0}^{k-1} t_i} \]

for

\[ D(x, s, \theta) = D_F(x, s, \theta) + D_f(x, s) \]

\[ = \frac{D_f(x + \theta(s - x), x)}{\theta} + D_\psi(x, s, \theta). \]

Curvature condition (cf. Jaggi’s curvature)

For some \( M > 0 \) and \( \nu > 0 \) and all \( x, s \in \text{dom}(\psi) \) and \( \theta \in [0, 1] \)

\[ D(x, s, \theta) \leq \frac{M \theta^\nu}{1 + \nu}. \]

This holds in particular when \( \text{dom}(\psi) \) bounded and \( \nabla f \) is \( \nu \)-Hölder continuous.
Theorem

If the above curvature condition holds and $\theta_k = \frac{1+\nu}{k+1+\nu}$ then

$$f(x_k) + \psi(x_k) + f^*(u_k) + \psi^*(-u_k) \leq M \left( \frac{1 + \nu}{k + 1 + \nu} \right)^\nu.$$ 

Proof: Let $\text{gap}(x_k, u_k) := f(x_k) + \psi(x_k) + f^*(u_k) + \psi^*(-u_k)$. Main Theorem implies that $\text{gap}(x_0, u_0) \leq D(x_0, s_0, 1)$ and

$$\text{gap}(x_{k+1}, u_{k+1}) \leq (1 - \theta_k)\text{gap}(x_k, u_k) + \theta_k D(x_k, s_k, \theta_k)$$

Curvature condition and induction show that

$$\text{gap}(x_k, u_k) \leq M \left( \frac{1 + \nu}{k + 1 + \nu} \right)^\nu.$$ 

The above generalizes the $O(1/k)$ convergence of conditional gradient.
Bregman proximal gradient

Want to solve \( \min_x \{ f(x) + \psi(x) \} \). Suppose \( f \) is differentiable.

Bregman proximal gradient

- pick \( s_{-1} \in \text{dom}(\psi) \)
- for \( k = 0, 1, \ldots \)
  - pick \( t_k > 0 \)
  - pick \( s_k \in \text{argmin}_s \left\{ \langle \nabla f(s_{k-1}), s \rangle + \psi(s) + \frac{1}{t_k} D_h(s, s_{k-1}) \right\} \)
- end for

This is the first-order meta-algorithm for

\[
y_k = s_{k-1}, \quad g_k = \nabla f(s_{k-1}).
\]
Recall Main Theorem

Let

\[ x_k := \frac{\sum_{i=0}^{k-1} t_i s_i}{\sum_{i=0}^{k-1} t_i}, \quad u_k := \frac{\sum_{i=0}^{k-1} t_i g_i}{\sum_{i=0}^{k-1} t_i}, \quad d_k(s) := \frac{D_h(s, s-1)}{\sum_{i=0}^{k-1} t_i}, \quad \theta_k := \frac{t_k}{\sum_{i=0}^{k} t_i}. \]

The iterates generated by the meta-algorithm satisfy

\[ f(x_k) + \psi(x_k) + f^*(u_k) + (\psi + d_k)^*(-u_k) \]
\[ \leq \frac{\sum_{i=0}^{k-1} (t_i (D_F(x_i, s_i, \theta_i) + D_f(s_i, y_i)) - D_h(s_i, s_{i-1}))}{\sum_{i=0}^{k-1} t_i} \]

for (recall \( F = f + \psi \))

\[ D_F(x, s, \theta) := \frac{F(x + \theta(s - x)) - (1 - \theta)F(x) - \theta F(s)}{\theta} \leq 0. \]

For notational convenience let \( x_0 := s_{-1} \) so that \( d_k(x) := \frac{D_h(x, x_{0})}{\sum_{i=0}^{k-1} t_i} \).
Theorem

Suppose the stepsizes satisfy \( t_i \cdot D_f(s_i, s_{i-1}) \leq D_h(s_i, s_{i-1}) \). Then for all \( x \in \mathbb{R}^n \)

\[
f(x_k) + \psi(x_k) - (f(x) + \psi(x)) \leq \frac{D_h(x, x_0)}{\sum_{i=0}^{k-1} t_i}
\]

Proof: Above condition on stepsizes and Main Theorem imply that

\[
f(x_k) + \psi(x_k) + f^*(u_k) + (\psi + d_k)^*(-u_k) \leq 0.
\]

Thus for all \( x \in \mathbb{R}^n \)

\[
f(x_k) + \psi(x_k) - (f(x) + \psi(x)) \leq d_k(x) = \frac{D_h(x, x_0)}{\sum_{i=0}^{k-1} t_i}.
\]
Smoothness and $\mathcal{O}(1/k)$ convergence of proximal gradient

Suppose $\bar{X} : = \arg\min_x \{ f(x) + \psi(x) \} \neq \emptyset$.

Relative smoothness

We say that $f$ is $L$-smooth relative to $h$ on $C$ if for all $x, y \in C$

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

It is easy to see that $f$ is $L$-smooth relative to $h$ if $\nabla f$ is $L$-Lipschitz continuous and $h(x) = \|x\|_2^2/2$

When $f$ is $L$-smooth relative to $h$ on $\text{dom}(\psi)$, we can guarantee

$$D_f(s_i, s_{i-1}) \leq \frac{1}{t_i} D_h(s_i, s_{i-1}) \text{ with } t_i \geq 1/L$$

and recover the iconic $\mathcal{O}(1/k)$ convergence rate for proximal gradient:

$$f(x_k) + \psi(x_k) - \min_x \{ f(x) + \psi(x) \} \leq \frac{L \cdot D_h(\bar{X}, x_0)}{k}. $$
Fast and universal Bregman proximal gradient

Fast and universal Bregman proximal gradient

- pick $x_0 := s_{-1} \in \text{dom}(\psi)$
- for $k = 0, 1, \ldots$
  - let $y_k := (1 - \theta_k)x_k + \theta_k s_{k-1}$
  - pick $t_k > 0$
  - pick $s_k \in \text{argmin}_s \left\{ \langle \nabla f(y_k), s \rangle + \psi(s) + \frac{1}{t_k} D_h(s, s_{k-1}) \right\}$
  - let $x_{k+1} := (1 - \theta_k)x_k + \theta_k s_k$
- end for

This is the first-order meta-algorithm for

$y_k = (1 - \theta_k)x_k + \theta_k s_{k-1}$, $g_k = \nabla f(y_k)$. 
Convergence of fast Bregman proximal gradient

Theorem

Suppose the stepsizes satisfy

\[ t_i \cdot (\mathcal{D}(x_i, s_i, \theta_i) + D_f(s_i, y_i)) \leq D_h(s_i, s_{i-1}). \]

Then for all \( x \in \mathbb{R}^n \)

\[ f(x_k) + \psi(x_k) - f(x) - \psi(x) \leq \frac{D_h(x, x_0)}{\sum_{i=0}^{k-1} t_i}. \]

Proof: Again condition on stepsizes and Main Theorem imply that

\[ f(x_k) + \psi(x_k) + f^*(u_k) + (\psi + d_k)^*(-u_k) \leq 0. \]

Thus for all \( x \in \mathbb{R}^n \)

\[ f(x_k) + \psi(x_k) - (f(x) + \psi(x)) \leq d_k(x) = \frac{D_h(x, x_0)}{\sum_{i=0}^{k-1} t_i}. \]
Triangle scaling and $\mathcal{O}(1/k^2)$ convergence

Triangle scaling (cf. Hanzely et al (2018))

Suppose for some $L > 0$ and all $x, s, s_\pm \in C$ and $\theta \in [0, 1]$

$$D_f((1 - \theta)x + \theta s, (1 - \theta)x + \theta s_\pm) \leq L \cdot \theta^2 \cdot D_h(s, s_\pm)$$

Observe

Triangle scaling $\implies$ Relative smoothness (take $\theta = 1$).
The converse holds when $h(x) = \|x\|_2^2/2$.

When triangle scaling condition holds, we can guarantee

$$t_i \cdot (D(x_i, s_i, \theta_i) + D_f(s_i, y_i)) \leq D_h(s_i, s_{i-1}) \text{ with } t_i \geq (i + 1)/L$$

and thus

$$f(x_k) + \psi(x_k) - \min_x \{f(x) + \psi(x)\} \leq \frac{2L \cdot D_h(\bar{X}, x_0)}{k(k+1)}.$$

Recover iconic $\mathcal{O}(1/k^2)$ convergence: Nesterov (1984), Beck-Teboulle (2009), Nesterov (2013), ...
Convergence of universal Bregman proximal gradient

Smoothness-plus condition

Suppose $\nu \in [0, 1]$ and $M > 0$ are such that for all $x, s, s_- \in C$ and $\theta \in [0, 1]$

$$D_f((1 - \theta)x + \theta s, (1 - \theta)x + \theta s_-) \leq \frac{2M\theta^{1+\nu}D_h(s, s_-)^{1+\nu}}{1 + \nu}.$$ 

Observe

Smoothness-plus holds if $h(x) = \|x\|^2_2/2$ and $\nabla f$ is $\nu$-Hölder continuous.
Convergence of universal Bregman proximal gradient

**Theorem**

Let \( \epsilon > 0 \) be fixed. Suppose the Smoothness-plus condition holds on \( \text{dom}(\psi) \) and \( t_i \) is the largest such that

\[
t_i \cdot (D(x_i, s_i, \theta_i) + D_f(s_i, y_i)) \leq D_h(s_i, s_{i-1}) + t_i \epsilon.
\]

Then for all \( x \in \mathbb{R}^n \)

\[
f(x_k) + \psi(x_k) - (f(x) + \psi(x)) \leq \frac{2M^{1+\nu} D_h(x, x_0)}{\epsilon^{1+\nu}} \frac{1}{k} \frac{1+3\nu}{1+\nu} + \epsilon.
\]

**Proof:** Main Theorem implies that

\[
f(x_k) + \psi(x_k) - f(x) - \psi(x) \leq d_k(x) + \epsilon = \frac{D_h(x, x_0)}{\sum_{i=0}^{k-1} t_i} + \epsilon.
\]

To finish: the Smoothness-plus condition yields

\[
\frac{1}{\sum_{i=0}^{k-1} t_i} \geq \frac{\theta_{k-1}}{t_{k-1}} \leq \frac{2M^{1+\nu}}{\epsilon^{1+\nu} k^{1+\nu}}.
\]

Recover \( O(1/k^{1+3\nu}/2) \) universal convergence by Nesterov (2015).
Stronger Convergence Results for Conditional Gradient
Conditional gradient revisited

Want to solve $\min_x \{f(x) + \psi(x)\}$. Suppose $f$ is differentiable and the mapping

$$g \mapsto \partial \psi^*(-g) = \arg\min \{\langle g, x \rangle + \psi(x)\}$$

is computable.

Conditional gradient

- pick $x_0 \in \text{dom}(f)$
- for $k = 0, 1, \ldots$
  - pick $s_k \in \arg\min_s \{\langle \nabla f(x_k), s \rangle + \psi(s)\}$ and $\theta_k \in [0, 1]$
  - let $x_{k+1} := (1 - \theta_k)x_k + \theta_ks_k$
- end for
Growth property

Recall

\[ \text{gap}(x, u) := f(x) + \psi(x) + f^*(u) + \psi^*(-u) \]

\[ D(x, s, \theta) := \frac{D_f(x + \theta(s - x), x)}{\theta} + \nabla \psi(x, s, \theta). \]

Observe: for \( x \in \text{dom}(\psi), g := \nabla f(x), \) and \( s \in \partial \psi^*(-g) \)

\[ \text{gap}(x, g) = \langle g, x - s \rangle + \psi(x) - \psi(s). \]

Growth property

Suppose \( \nu > 0 \) and \( r \in [0, 1] \). Say that \((D, \text{gap})\) satisfies the \((\nu, r)\)-growth property if there exists \( M > 0 \) such that for all \( x \in \text{dom}(\psi), g := \nabla f(x), \) and \( s \in \partial \psi^*(-g) \)

\[ D(x, s, \theta) \leq \frac{M \theta^\nu}{1 + \nu} \cdot \text{gap}(x, g)^r \text{ for all } \theta \in [0, 1]. \]
Growth property: special cases

Case $r = 0$
In this case the growth property is

$$D(x, s, \theta) \leq \frac{M\theta^\nu}{1 + \nu} \text{ for all } \theta \in [0, 1].$$

This is the same as the *curvature condition* discussed earlier. It holds if $\nabla f$ is $\nu$-Hölder continuous and $\text{dom}(\psi)$ is bounded.

Case $\nu = 1$ and $r = 1$
In this case the growth property is

$$D(x, s, \theta) \leq \frac{M\theta}{2} \cdot \text{gap}(x, g) \text{ for all } \theta \in [0, 1].$$

It holds if $\nabla f$ is Lipschitz continuous and $\psi$ is strongly convex.

Other cases with $\nu > 0$, $r \in (0, 1)$ when $f$ is uniformly smooth and $\psi$ is uniformly convex.
Let $x_0, x_1, \ldots$ denote the iterates generated by the conditional gradient algorithm. For $k = 0, 1, \ldots$ let

$$\text{bestgap}_k := \min_{i=0,1,\ldots,k} \text{gap}(x_k, g_i)$$

where $g_i = \nabla f(x_i)$ for $i = 0, 1, \ldots$.

**Line-search procedure**

Choose $\theta_k \in [0, 1]$ via

$$\theta_k := \arg\min_{\theta \in [0,1]} \{(1 - \theta) \cdot \text{gap}(x_k, g_k) + \theta \cdot D(x_k, s_k, \theta)\}.$$
Growth property and convergence rates

**Theorem**

Suppose \((D, \text{gap})\) satisfy the \((\nu, r)\)-growth and \(\theta_k\) is as above. For \(r = 1\) we have linear convergence

\[
\text{bestgap}_k \leq \text{bestgap}_0 \left(1 - \frac{\nu}{\nu + 1} \cdot \frac{1}{M^{\frac{1}{\nu}}} \right)^k.
\]

For \(r \in [0, 1)\) we have an initial linear convergence regime

\[
\text{bestgap}_k \leq \text{bestgap}_0 \left(1 - \frac{\nu}{\nu + 1}\right)^k, \quad k = 0, 1, 2, \ldots, k_0
\]

where \(k_0\) is the smallest \(k\) such that \(\text{bestgap}_k^{1-r} \leq M\). Then for \(k \geq k_0\) we have a sublinear convergence regime

\[
\text{bestgap}_k \leq \left(\text{bestgap}_{k_0}^{\frac{r-1}{\nu}} + \frac{1-r}{\nu + 1} \cdot \frac{1}{M^{\frac{1}{\nu}}} \cdot (k - k_0)\right)^{\frac{\nu}{r-1}}.
\]
Conclusions

Consider the problem \( \min_{x \in \mathbb{R}^n} \{ f(x) + \psi(x) \} \) where \( f, \psi \) convex.

- Perturbed Fenchel duality: first-order meta-algorithm generates iterates that satisfy

\[
f(x_k) + \psi(x_k) + f^*(u_k) + (\psi + d_k)^*(-u_k) \leq \delta_k
\]

- Convergence of popular first-order methods readily follow:
  - \( O(1/k^\nu) \) for conditional gradient if curvature condition holds
  - \( O(1/k) \) for proximal gradient if relative smoothness holds
  - \( O(1/k^2) \) for fast proximal gradient if triangle scaling holds
  - \( O(1/\sqrt{k}) \) for subgradient if relative continuity holds (skipped)

- Stronger convergence rates for conditional gradient if some suitable growth property holds.

- Above holds for more general problem \( \min_{x \in \mathbb{R}^n} \{ f(Ax) + \psi(x) \} \) and its dual \( \max_{u \in \mathbb{R}^n} \{ -f^*(u) - \psi^*(-A^*u) \} \).
Main references

- Gutman and P. “Perturbed Fenchel duality and first-order methods,” *Mathematical Programming*.

- P. “Affine invariant convergence rates of the conditional gradient method,” https://arxiv.org/abs/2112.06727
Uniform smoothness and uniform convexity

Let \( q \in (1, 2] \). Say that \( f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\} \) is \( q \)-uniformly smooth if there exist \( L > 0 \) such that for all \( x, y \in \mathbb{R}^n \) and \( \theta \in [0, 1] \)

\[
f(x + \theta(y - x)) \geq (1 - \theta)f(x) + \theta f(y) - \frac{L}{q} \theta(1 - \theta)\|y - x\|^q.
\]

Let \( p \geq 2 \). Say that \( \psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\} \) is \( p \)-uniformly convex if there exist \( \mu > 0 \) such that for all \( x, y \in \mathbb{R}^n \) and \( \theta \in [0, 1] \)

\[
\psi(x + \theta(y - x)) \leq (1 - \theta)\psi(x) + \theta \psi(y) - \frac{\mu}{p} \theta(1 - \theta)\|y - x\|^p.
\]

Facts

- If \( f \) is \( q \)-unif smooth and \( \psi \) is \( p \)-unif convex then \((D, \text{gap})\) satisfies the \((\nu, r)\)-growth property for \( \nu = q - 1 \) and \( r = q/p \).
- \( f \) is \((\nu + 1)\)-uniformly smooth if \( \nabla f \) is \( \nu \)-Hölder continuous.
- \( f \) is \( q \)-unif smooth iff \( f^* \) is \( p \)-unif convex for \( 1/p + 1/q = 1 \).