Perturbed Fenchel Duality and First-Order Methods

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

WOMBAT 2021, December 2021

Preamble: some motivation

Convex optimization

Constrained format

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

where $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $C \subseteq \mathbb{R}^n$ are convex and C has some "simple" structure.

Composite minimization format

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

where $f, \psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ are convex and ψ has some "simple" structure.

Composite format subsumes the constrained format 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 $\Pi_C : \mathbb{R}^n \to C$ denote the orthogonal projection onto C.

Projected subgradient method (SG)

$$\begin{aligned} &\text{pick } g_k \in \partial f(x_k) \ \text{ and } \ t_k > 0 \\ &x_{k+1} = \Pi_C(x_k - t_k g_k) \end{aligned}$$

Projected gradient descent (GD)

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

Conditional gradient (CG)

$$\begin{aligned} &s_k = \operatorname*{argmin}_{s \in C} \langle \nabla f(x_k), s \rangle \\ &\text{pick } \ \theta_k \in [0, 1] \\ &x_{k+1} = x_k + \theta_k (s_k - x_k) \end{aligned}$$

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

Suppose the following proximal mapping is computable for all $t > 0$

$$g \mapsto \mathsf{Prox}_t(g) := \underset{y\in\mathbb{R}^n}{\mathsf{argmin}} \left\{\psi(y) + \frac{1}{2t}\|y-g\|^2\right\}$$

Observe: if $\psi = \delta_C$ then $\mathsf{Prox}_t = \Pi_C$ for all $t > 0$.

Proximal gradient (PG)

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

Fast proximal gradient (FPG)

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

(Nesterov (1984), Beck-Teboulle (2009), Nesterov (2013),...)

Convergence properties

Under suitable assumptions of smoothness and choice of stepsizes:

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

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

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 ϵ-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

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

for some $\delta_k \geq 0$ 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

$$f(x_k) + \psi(x_k) - (f(x) + \psi(x)) \leq \delta_k + d_k(x).$$

First-Order Meta-Algorithm

First-order meta-algorithm

Want to solve $\min_x \{f(x) + \psi(x)\}$.
Suppose the following proximal mapping is computable for all $t > 0$

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

Key ideas

- Generate two sequences s_k, y_k
- At iteration k pick $g_k \in \partial f(y_k)$ and $t_k > 0$ and update s_k via

$$s_k = \mathsf{Prox}_{t_k}(s_{k-1} - t_k g_k)$$

- Flexibility on the selection of y_k.

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

First-order meta-algorithm

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

First-order meta-algorithm

```
• pick $s_{-1} \in \mathrm{dom}(\psi)$
• for $k = 0, 1, \ldots$
    pick $y_k \in \mathrm{dom}(\partial f),\ g_k \in \partial f(y_k)$, and $t_k > 0$
    let $s_k := \mathsf{Prox}_{t_k}(s_{k-1} - t_k g_k)$
  end for
```

Some convenient notation

Let $F := f + \psi$ and for $g \in \partial f(y)$ let $D_f(x, y)$ denote the following *Bregman distance*

$$D_f(x, y) := f(x) - f(y) - \langle g, x - y \rangle.$$

Main Theorem

Let $x_0 := s_{-1}$ and

$$x_k = \frac{\sum_{i=0}^{k-1} t_i s_i}{\sum_{i=0}^{k-1} t_i},\ u_k = \frac{\sum_{i=0}^{k-1} t_i g_i}{\sum_{i=0}^{k-1} t_i},\ d_k(s) = \frac{\|s - x_0\|^2}{2\sum_{i=0}^{k-1} t_i},\ \theta_k = \frac{t_k}{\sum_{i=0}^{k} t_i}.$$

Theorem

The iterates generated by the above meta-algorithm satisfy

$$\begin{aligned} & f(x_k) + \psi(x_k) + f^*(u_k) + (\psi + d_k)^*(-u_k) \\ & \qquad \le \frac{\sum_{i=0}^{k-1} \left(t_i \mathcal{D}(x_i, y_i, s_i, \theta_i)/\theta_i - \|s_i - s_{i-1}\|^2/2\right)}{\sum_{i=0}^{k-1} t_i} \end{aligned}$$

for

$$\mathcal{D}(x, y, s, \theta) := F(x + \theta(s - x)) - (1 - \theta)F(x) - \theta F(s) + \theta D_f(s, y).$$

Convergence of Iconic First-Order Algorithms

Proximal gradient

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

Proximal gradient

- pick $y_0 \in \mathrm{dom}(\psi)$
- for $k = 0, 1, \ldots$

```
    pick $t_k > 0$
    let $y_{k+1} = \mathsf{Prox}_{t_k}(y_k - t_k \nabla f(y_k))$
end for
```

This is precisely the first-order meta-algorithm with $y_k = s_{k-1}$.

Proximal gradient

Recall $F = f + \psi$ and

$$\begin{aligned}\mathcal{D}(x,y,s,\theta) &= F(x+\theta(s-x)) - (1-\theta)F(x) - \theta F(s) + \theta D_f(s,y) \\ &\leq \theta D_f(s,y).\end{aligned}$$

Thus Main Theorem yields

$$\begin{aligned}&f(x_k) + \psi(x_k) + f^*(u_k) + (\psi + d_k)^*(-u_k) \\ &\qquad\leq \frac{\sum_{i=0}^{k-1}\big(t_i\mathcal{D}(x_i,y_i,s_i,\theta_i)/\theta_i - \|s_i - s_{i-1}\|^2/2\big)}{\sum_{i=0}^{k-1} t_i} \\ &\qquad\leq \frac{\sum_{i=0}^{k-1}\big(t_i D_f(s_i,s_{i-1}) - \|s_i - s_{i-1}\|^2/2\big)}{\sum_{i=0}^{k-1} t_i}.\end{aligned}$$

Theorem

Suppose the stepsizes satisfy $D_f(s_i, s_{i-1}) \leq \frac{1}{2t_i}\|s_i - s_{i-1}\|^2$. Then for all $x \in \mathbb{R}^n$

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

Proof: Main Theorem implies that

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

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

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

□

Smoothness and $\mathcal{O}(1/k)$ convergence of proximal gradient

Suppose $\bar{X} := \mathsf{argmin}_x\{f(x) + \psi(x)\} \neq \emptyset$.

Smoothness

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

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

It is easy to see that f is L-smooth if ∇f is L-Lipschitz.

When f is L-smooth on $\mathrm{dom}(\psi)$, we can take $t_i \geq 1/L$ and recover the iconic $\mathcal{O}(1/k)$ convergence rate for proximal gradient:

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

Fast and universal proximal gradient

Fast and universal proximal gradient

```
• pick $x_0 := s_{-1} \in \mathrm{dom}(\psi)$
• for $k = 0, 1, \ldots$
      let $y_k := (1 - \theta_k)x_k + \theta_k s_{k-1}$ and pick $t_k > 0$
      let $s_k := \mathrm{Prox}_{t_k}(s_{k-1} - t_k \nabla f(y_k))$
      let $x_{k+1} := (1 - \theta_k)x_k + \theta_k s_k$
  end for
```

First-order meta-algorithm with $y_k = (1 - \theta_k)x_k + \theta_k s_{k-1}$.

Observe: the sequence y_k can also be written as

$$y_k = x_k + \frac{\theta_k(1 - \theta_{k-1})}{\theta_{k-1}}(x_k - x_{k-1})$$

Convergence of fast proximal gradient

Theorem

Suppose t_i *is such that* $t_i \cdot \mathcal{D}(x_i, y_i, s_i, \theta_i)/\theta_i \leq \|s_i - s_{i-1}\|^2/2.$
Then for all $x \in \mathbb{R}^n$

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

Proof: Again Main Theorem implies that

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

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

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

□

Smoothness and $\mathcal{O}(1/k^2)$ convergence

Recall that f is L-smooth on C if for all $x, y \in C$

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

Fast proximal gradient: when f is L-smooth on $\mathrm{dom}(\psi)$ we have

$$\mathcal{D}(x_i, y_i, s_i, \theta_i) \le \frac{L \cdot \theta_i^2 \|s_i - s_{i-1}\|^2}{2}.$$

Thus we can take t_i such that $t_i \theta_i \ge 1/L$. This implies that

$$\frac{1}{\sum_{i=0}^{k-1} t_i} = \frac{\theta_{k-1}}{t_{k-1}} \le L \left(\frac{2}{k+1} \right)^2.$$

Recover iconic $\mathcal{O}(1/k^2)$ convergence for fast proximal gradient:

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

Nesterov (1984), Beck-Teboulle (2009), Nesterov (2013), ...

Convergence of universal proximal gradient

Smoothness-like condition

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

$$D_f(x,y) \leq \frac{M\|x-y\|^{1+\nu}}{1+\nu}.$$

Observe

Smothness-like holds if ∇f is ν-Hölder continuous.

Convergence of universal proximal gradient

Theorem

Let $\epsilon > 0$ be fixed. Suppose the smoothness-like condition holds on $\operatorname{dom}(\psi)$ and t_i is the largest such that

$$t_i \cdot \mathcal{D}(x_i, y_i, s_i, \theta_i)/\theta_i \le \|s_i - s_{i-1}\|^2/2 + t_i\epsilon.$$

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

$$f(x_k) + \psi(x_k) - (f(x) + \psi(x)) \le \frac{M^{\frac{2}{1+\nu}}\|x - x_0\|^2}{\epsilon^{\frac{1-\nu}{1+\nu}} k^{\frac{1+3\nu}{1+\nu}}} + \epsilon.$$

Proof: Main Theorem implies that

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

To finish: the smoothness-like condition yields

$$\frac{1}{\sum_{i=0}^{k-1} t_i} = \frac{\theta_{k-1}}{t_{k-1}} \le \frac{2M^{\frac{2}{1+\nu}}}{\epsilon^{\frac{1-\nu}{1+\nu}} k^{\frac{1+3\nu}{1+\nu}}}. \quad \square$$

Recover $\mathcal{O}(1/k^{\frac{1+3\nu}{2}})$ universal convergence by Nesterov (2015).

First-Order Meta-Algorithm (non-Euclidean)

First-order meta-algorithm (non-Euclidean)

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 $t > 0$:

$$(g, s_-) \mapsto \operatorname*{argmin}_s \left\{ \langle g, s \rangle + \psi(s) + \frac{1}{t} D_h(s, s_-) \right\}.$$

Example

$h(x) = \|x\|_2^2/2 \rightsquigarrow D_h(y, x) = \|y - x\|_2^2/2.$

First-order meta-algorithm (non-Euclidean)

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

First-order meta-algorithm (non-Euclidean)

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

Why consider non-Euclidean algorithms?

- 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 mapping

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

is easily computable for $h(x) = \sum_{i=1}^n x_i \log(x_i)$.

Main Theorem again

Let

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

Theorem

The iterates generated by the above meta-algorithm satisfy

$$\begin{aligned} f(x_k) + \psi(x_k) + f^*(u_k) + (\psi + d_k)^*(-u_k) \\ \leq \frac{\sum_{i=0}^{k-1} \left(t_i \mathcal{D}(x_i, y_i, s_i, \theta_i)/\theta_i - D_h(s_i, s_{i-1})\right)}{\sum_{i=0}^{k-1} t_i} \end{aligned}$$

for

$$\mathcal{D}(x, y, s, \theta) := F(x + \theta(s - x)) - (1 - \theta)F(x) - \theta F(s) + \theta D_f(s, y).$$

(Recall that $F = f + \psi$.)

Some special cases of first-order meta-algorithm

Bregman proximal gradient

Obtained by taking $y_k = s_{k-1}$. Get $\mathcal{O}(1/k)$ convergence if the following *relative* L-smoothness assumption holds:

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

This $\mathcal{O}(1/k)$ convergence result was established by Bauschke et al. (2016) and by Lu et al. (2018).

Fast and universal Bregman proximal gradient

Obtained by taking $y_k = (1-\theta_k)x_k + \theta_k s_{k-1}$. Get $\mathcal{O}(1/k^{\frac{1+3\nu}{2}})$ convergence if the following smoothness-like property holds:

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

Related *triangle-scaling* property by Hanzely et al (2018).

Two more special cases: conditional gradient and proximal subgradient.

Conditional gradient

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

Conditional gradient

- pick $x_0 \in \operatorname{dom}(f)$
- for $k = 0, 1, \ldots$

```
    let $g_k := \nabla f(x_k)$
    pick $s_k \in \operatorname{argmin}_s \{\langle g_k, s\rangle + \psi(s)\}$ and $\theta_k \in [0,1]$
    let $x_{k+1} := (1-\theta_k)x_k + \theta_k s_k$
end for
```

This is the first-order meta-algorithm for

$$s_{-1} = x_0,\ y_k = x_k,\ h \equiv 0,\ \text{and } t_k \text{ such that } \ \theta_k = \frac{t_k}{\sum_{i=1}^{k} t_i}.$$

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

Conditional gradient

For the conditional gradient algorithm the Main Theorem yields

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

for

$$\begin{aligned} D(x, s, \theta) = D_f(x + \theta(s - x), x) \\ + \psi(x + \theta(s - x)) - (1 - \theta)\psi(x) - \theta\psi(s). \end{aligned}$$

Curvature condition (cf. Jaggi's curvature)

For $\nu > 0$ there exists $M > 0$ such that for all $x, s \in \mathrm{dom}(\psi)$ and $\theta \in [0, 1]$

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

This holds in particular when $\mathrm{dom}(\psi)$ bounded and ∇f is ν-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: Main Theorem implies that

$$f(x_k) + \psi(x_k) + f^*(u_k) + \psi^*(-u_k) \leq \mathsf{CGgap}_k$$

where $\mathsf{CGgap}_0 = D(x_0, s_0, 1)$ and

$$\mathsf{CGgap}_{k+1} = (1-\theta_k)\mathsf{CGgap}_k + D(x_k, s_k, \theta_k), \; k = 1, 2, \ldots.$$

Curvature condition and induction show that

$$\mathsf{CGgap}_k \leq M \left(\frac{1+\nu}{k+1+\nu} \right)^{\nu}.$$

□

The above generalizes the $\mathcal{O}(1/k)$ convergence of conditional gradient.

Sublinear to linear spectrum of convergence rates

Define the duality gap function

$$\mathsf{gap}(x,u) := f(x) + \psi(x) + f^*(u) + \psi^*(-u).$$

Curvature-like condition

For $\nu > 0$ and $r \in [0,1]$ there exists $M \geq 1$ such that for $x \in \mathrm{dom}(\psi),\ g := \nabla f(x),\ s = \mathsf{argmin}_y\{\langle g, y\rangle + \psi(y)\}$ and $\theta \in [0,1]$

$$D(x,s,\theta) \leq \frac{M\theta^{1+\nu}}{1+\nu} \cdot \mathsf{gap}(x,g)^r.$$

Previous curvature condition corresponds to special case $r = 0$. Special case $\nu = 1, r = 1$ holds when ∇f is Lipschitz continuous and ψ is strongly continuous.

Line-search procedure

Choose $\theta_k \in [0,1]$ in the conditional gradient algorithm via

$$\theta_k := \mathop{\mathsf{argmin}}_{\theta \in [0,1]}\{(1-\theta)\mathsf{gap}(x_k, g_k) + D(x_k, s_k, \theta)\}.$$

Sublinear to linear spectrum of convergence rates

Consider the "best duality gap": $\mathsf{gap}_k := \min_{i=0,1,\ldots,k} \mathsf{gap}(x_k, g_k)$.

Theorem

Suppose the curvature-like condition holds for some $\nu > 0, r \in [0,1]$ and the conditional gradient algorithm chooses $\theta_k \in [0,1]$ via above line-search procedure.
If $r = 1$ then $\mathsf{gap}_k \to 0$ linearly:

$$\mathsf{gap}_k \leq \left(1 - \frac{\nu}{(\nu+1)M^{\frac{1}{\nu}}}\right)^k \mathsf{gap}_0.$$

If $r \in [0,1)$ then for $k \leq k_0 := \mathrm{argmin}\{i : \mathsf{gap}_i \leq 1\}$

$$\mathsf{gap}_k \leq \left(1 - \frac{\nu}{\nu+1}\right)^k \mathsf{gap}_0,$$

and for $k > k_0$

$$\mathsf{gap}_k \leq \left(\mathsf{gap}_{k_0}^{\frac{r-1}{\nu}} + \frac{1-r}{(\nu+1)M^{\frac{1}{\nu}}}(k-k_0)\right)^{\frac{\nu}{r-1}}.$$

Conclusions

Consider the problem $\min_{x\in\mathbb{R}^n} \{f(x)+\psi(x)\}$ where f,ψ 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 most popular first-order methods readily follows.

Main references

Gutman and P. *"Perturbed Fenchel duality and first-order methods,"* https://arxiv.org/abs/1812.10198

P. *"Affine-invariant convergence rates of the conditional gradient method".* Forthcoming.