Proximal Gradient Algorithm

Andersen Ang

Department of Combinatorics and Optimization,
University of Waterloo, Waterloo, Canada

ms$\mathbf{x}$ang@uwaterloo.ca, where $\mathbf{x} = \lfloor \pi \rfloor$
Homepage: angms.science

First draft: August 2, 2017 Last update: July 9, 2021

Overview

1. Constrained optimization, projected gradient and proximal gradient
2. Proximal Gradient Descent converges at order $\mathcal{O}(\frac{1}{k})$
3. Summary

Constrained optimization and projected gradient

- Constrained minimization $\min_{\mathbf{x}\in\mathcal{C}\subset\mathbb{R}^n} f(\mathbf{x})$ with $\mathcal{C}\subseteq\mathbb{R}^n$ is a closed compact convex set.

- Starting from an initial point $\mathbf{x}_0\in\mathcal{C}$, Projected Gradient Descent (PGD) iterates

$$\mathbf{x}_{k+1} = \mathrm{proj}_{\mathcal{C}}\Big(\mathbf{x}_k - t_k\nabla f(\mathbf{x}_k)\Big),$$

  where k is the iteration counter, t_k is the stepsize and $\mathrm{proj}_{\mathcal{C}}(\,\cdot\,)$ is the projection operator

$$\mathrm{proj}_{\mathcal{C}}(\mathbf{x}_0) = \arg\min_{\mathbf{x}\in\mathcal{C}} \frac{1}{2}\|\mathbf{x}-\mathbf{x}_0\|_2^2,$$

  i.e. given a point $\mathbf{x}_0$, $\mathrm{proj}_{\mathcal{C}}$ find a point $\mathbf{x}\in\mathcal{C}$ which is "closest" to $\mathbf{x}_0$.

Proximal operator $\mathrm{prox}_{\alpha f}(\mathbf{x}_0)$

- Given a function f, a parameter $\alpha \geq 0$ and a point $\mathbf{x}_0$,

$$\mathrm{prox}_{\alpha f}(\mathbf{x}_0) \coloneqq \underset{\mathbf{x}}{\mathrm{argmin}} \left\{ f(\mathbf{x}) + \frac{1}{2\alpha} \|\mathbf{x} - \mathbf{x}_0\|_2^2 \right\}.$$

  - $\mathrm{prox}_{\alpha f}(\mathbf{x}_0) =$ find a $\mathbf{x}$ that "compromises" between the distance from $\mathbf{x}_0$ and the minimization of f

- Compared with $\mathrm{proj}_{\mathcal{C}}(\mathbf{x}_0) = \arg\min_{\mathbf{x} \in \mathcal{C}} \frac{1}{2} \|\mathbf{x} - \mathbf{x}_0\|_2^2$
  - Both are argmin functions that return a point
  - Notice that $\mathrm{proj}_{\mathcal{C}}(\mathbf{x}_0) = \mathrm{prox}_{\alpha f}(\mathbf{x}_0)$ with $f(\mathbf{x}) = 0,\ \alpha = 1$, which also means that $\mathrm{proj}_{\mathcal{C}}(\,\cdot\,)$ is the special case of $\mathrm{prox}_{\alpha f}(\,\cdot\,)$ if f is the indicator function of $\mathcal{C}$:

$$f(\mathbf{x}) = \begin{cases} 0 & \mathbf{x} \in \mathcal{C} \\ \infty & \mathbf{x} \notin \mathcal{C} \end{cases}.$$

On the parameter α

- In fact, by definition we have $\text{prox}_g(\mathbf{x}_0) \coloneqq \underset{\mathbf{x}}{\text{argmin}}\left\{ g(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2 \right\}$. So if $g = \alpha f$, we have

$$\begin{aligned}
\text{prox}_{\alpha f}(\mathbf{x}_0) &= \underset{\mathbf{x}}{\text{argmin}}\left\{ \alpha f(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2 \right\} \\
&= \underset{\mathbf{x}}{\text{argmin}}\left\{ \alpha\left(f(\mathbf{x}) + \frac{1}{2\alpha}\|\mathbf{x} - \mathbf{x}_0\|_2^2 \right) \right\} \\
&= \underset{\mathbf{x}}{\text{argmin}}\left\{ f(\mathbf{x}) + \frac{1}{2\alpha}\|\mathbf{x} - \mathbf{x}_0\|_2^2 \right\}.
\end{aligned}$$

- α controls the degree of compromises: larger α makes the second term less important

- For a sufficient small α, $f(\mathbf{x}) + \frac{1}{2\alpha}\|\mathbf{x} - \mathbf{x}_0\|_2^2$ is convex even if f is nonconvex

Comparing PGD and Proximal GD

- Problem $\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x})$ can be converted to an unconstrained problem with augmented cost function as $\min_{\mathbf{x}} f(\mathbf{x}) + i_{\mathcal{C}}(\mathbf{x})$, where $\mathbf{x} \in \mathcal{C}$ is expressed by the indicator function $i_{\mathcal{C}}(\mathbf{x})$.

- To solve $\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x})$, PGD step is $\mathbf{x}_{k+1} = \mathrm{proj}_{\mathcal{C}}\Big(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)\Big)$.

- To solve $\min_{x} f(\mathbf{x}) + i_{\mathcal{C}}(\mathbf{x})$, the Proximal GD step is $\mathbf{x}_{k+1} = \mathrm{prox}_{t_k i_{\mathcal{C}}}\Big(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)\Big)$, which is equivalent to $\mathbf{x}_{k+1} = \mathrm{prox}_{i_{\mathcal{C}}}\Big(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)\Big)$

- Both PGD and Proximal GD iterations are "forward-backward" iteration, forward refers to the gradient step $\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)$ and backward refers to proj and prox.

- Normally $\mathcal{C}$ and $i_{\mathcal{C}}$ are assumed to be "simple" (e.g. convex, algebraic) so that proj and prox are easy and cheap/fast to compute.

Example 1. $\min f(\mathbf{x}) + \|\mathbf{x}\|_1$, f differentiable

- The Proximal GD update is

$$\begin{aligned}
\mathbf{x}_{k+1} &= \operatorname{prox}_{t_k\|\mathbf{x}\|_1}\Big(\mathbf{x}_k - t_k\nabla f(\mathbf{x}_k)\Big) \\
&= \arg\min_{\mathbf{x}}\Big\{\|\mathbf{x}\|_1 + \frac{1}{2t_k}\|\mathbf{x} - \mathbf{x}_k + t_k\nabla f(\mathbf{x}_k)\|_2^2\Big\} \\
&= \arg\min_{\mathbf{x}}\Big\{\sum_i^n\Big(|x_i| + \frac{1}{2t_k}(x_i - [\mathbf{x}_k]_i + t_k[\nabla f(\mathbf{x}_k)]_i)^2\Big)\Big\}.
\end{aligned}$$

- Which becomes the element-wise soft-thresholding

$$\mathbf{x}_{k+1} = S_{t_k}\Big(\mathbf{x}_k - t_k\nabla f(\mathbf{x}_k)\Big), \quad \text{where} \quad [S_\alpha(\mathbf{u})]_i = \begin{cases} u_i - \alpha & u_i \geq \alpha \\ 0 & -\alpha \leq u_i \leq \alpha \\ u_i + \alpha & u_i \leq -\alpha \end{cases}$$

Comparing PGD and Proximal GD - on algorithm

PGD (for $\min_{\mathbf{x}\in\mathcal{C}} f(\mathbf{x})$)

- Start from an initial point $\mathbf{x}_0 \in \mathcal{C}$, loop the following until stopping condition is met
  1. Descent direction: pick the descent direction as $-\nabla f(\mathbf{x}_k)$
  2. Step size: pick a step size t_k
  3. Update: $\mathbf{y}_{k+1} = \mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)$
  4. Projection: $\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}\in\mathcal{C}} \frac{1}{2}\|\mathbf{x} - \mathbf{y}_{k+1}\|_2^2$

Proximal GD (for $\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x})$ where g is the indicator function of $\mathcal{C}$)

- Start from an initial point $\mathbf{x}_0 \in \mathcal{C}$, loop the following until stopping condition is met
  1. Descent direction: pick the descent direction as $-\nabla f(\mathbf{x}_k)$
  2. Step size: pick a step size t_k
  3. Update: $\mathbf{y}_{k+1} = \mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)$
  4. Solving optimization: $\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} \left\{ g(\mathbf{x}) + \frac{1}{2t_k}\|\mathbf{x} - \mathbf{y}_{k+1}\|_2^2 \right\}$

General problem setting for Proximal GD

- General problem

$$\min_{\mathbf{x}} F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x}),$$

  - f is convex, differentiable (possibly L-smooth, m-strongly convex)
  - g is convex (possibly non-differentiable), $\mathrm{prox}_g(\mathbf{x})$ is inexpensive and easy to compute
  - F^* is finite and exists (at minimizer $\mathbf{x}^*$, possibly non-unique)

- Proximal gradient update

$$\mathbf{x}_{k+1} = \mathrm{prox}_{t_k g}\Big(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)\Big)$$

  - the solution to $\mathrm{prox}_{y_k g}\big(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)\big)$ exists[1]
  - the update provides an unique solution $\mathbf{x}_{k+1}$, if $f + g$ is convex

[1] g has to be proper closed function and $g(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2$ is coerccive.

Gradient mapping of Proximal GD update

$$
\begin{aligned}
\mathbf{x}_{k+1} &= \operatorname{prox}_{t_k g}\Big(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)\Big) \\
&\overset{\text{tricky step}}{=} \mathbf{x}_k - \Big\{\mathbf{x}_k - \operatorname{prox}_{\alpha g}\Big(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)\Big)\Big\} \\
&= \mathbf{x}_k - t_k \frac{1}{t_k}\Big\{\mathbf{x}_k - \operatorname{prox}_{\alpha g}\Big(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)\Big)\Big\} \\
&= \mathbf{x}_k - t_k G_{t_k}(\mathbf{x}_k)
\end{aligned}
$$

where

$$
G_t(\mathbf{x}) = \frac{\mathbf{x} - \operatorname{prox}_{tg}\Big(\mathbf{x} - t\nabla f(\mathbf{x})\Big)}{t}.
$$

- The expression $G_{t_k}(\mathbf{x}_k)$ is called the gradient map of Proximal GD.
- $G_t(\mathbf{x})$ is not the gradient nor subgradient of $f(\mathbf{x}) + g(\mathbf{x})$
- $G_t(\mathbf{x}^*) = \mathbf{0}$ if $\mathbf{x}^*$ is the minimizer $f(\mathbf{x}^*) + g(\mathbf{x}^*)$

Example 2. $\min f(\mathbf{x}) + \|\mathbf{x}\|_2^2$, f differentiable

- The Proximal GD step (prox on $\|\mathbf{x}\|_2^2$) is

$$\mathbf{x}_{k+1} = \mathrm{prox}_{t_k\|\mathbf{x}\|_2^2}\Big(\mathbf{x}_k - t_k\nabla f(\mathbf{x}_k)\Big) = \underset{\mathbf{x}}{\mathrm{argmin}} \underbrace{\|\mathbf{x}\|_2^2 + \frac{1}{2t_k}\|\mathbf{x} - \mathbf{x}_k - t_k\nabla f(\mathbf{x}_k)\|_2^2}_{F}.$$

- Note that F is sum of two convex parts so it is convex. To minimize F, set $\nabla F = \mathbf{0}$ gives

$$\mathbf{x}_{k+1} = \frac{\mathbf{x}_k - t_k\nabla f(\mathbf{x}_k)}{2t_k + 1}.$$

- Note: in general, if $g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x} + \mathbf{b}^\top\mathbf{x} + c$ for $\mathbf{A} \in \mathbb{S}_+^n, \mathbf{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$, then

$$\mathrm{prox}_g(\mathbf{x}_0) = (\mathbf{A} + \mathbf{I})^{-1}(\mathbf{x}_0 - \mathbf{b})$$

  i.e., $\|\mathbf{x}\|_2^2$ is th special case with $\mathbf{A} = 2\mathbf{I}, \mathbf{b} = \mathbf{0}, c = 0$.

Iteration $\frac{1}{2t_k+1}\big(\mathbf{x}_k - t_k\nabla f(\mathbf{x}_k)\big)$ does solve $\min f(\mathbf{x}) + \|\mathbf{x}\|_2^2$

- Solving $\min f(\mathbf{x}) + \|\mathbf{x}\|_2^2$ by the update $\mathbf{x}_{k+1} = \dfrac{1}{2t_k+1}\big(\mathbf{x}_k - t_k\nabla f(\mathbf{x}_k)\big)$ seems wired.

- For $\min f(\mathbf{x}) + \|\mathbf{x}\|_2^2$, as f is differentiable, by 1st-order optimality condition

$$\nabla f(\mathbf{x}^*) + 2\mathbf{x}^* = \mathbf{0}.$$

- Rewriting the proximal gradient update as a gradient map gives

$$G(\mathbf{x}^*) = \frac{\mathbf{x}^* - \mathrm{prox}_{t\|\mathbf{x}\|_2^2}\Big(\mathbf{x}^* - t\nabla f(\mathbf{x}^*)\Big)}{t} = \frac{\mathbf{x}^* - \dfrac{\mathbf{x}^* - t\nabla f(\mathbf{x}^*)}{2t+1}}{t}.$$

So $G(\mathbf{x}^*) = \mathbf{0}$ gives $\mathbf{x}^* - \dfrac{\mathbf{x}^* - t\nabla f(\mathbf{x}^*)}{2t+1} = \mathbf{0}$, which also gives $\nabla f(\mathbf{x}^*) + 2\alpha\mathbf{x}^* = \mathbf{0}$.

1 page review on sub-gradient

- For a differentiable convex f, we have

$$f(\mathbf{y}) \geq \underbrace{f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})}_{\phi(\mathbf{x})}, \text{ for any } \mathbf{x}, \mathbf{y}$$

  $\phi(\mathbf{x})$ is the 1st-order Taylor approximation of f at $\mathbf{x}$, it forms a global underestimator of f at the point $\mathbf{x}$ for all $\mathbf{y}$.

- Such idea can be extended to non-differentiable convex h: for any $\mathbf{x}, \mathbf{y}$ we have

$$h(\mathbf{y}) \geq h(\mathbf{x}) + \mathbf{g}^\top (\mathbf{y} - \mathbf{x}),$$

  then the vector $\mathbf{g}$ is called *subgradient*
  - subgradient always exists
  - unlike gradient, function can have multiple subgradients at a point
  - the set of all subgradients at $\mathbf{x}$ is the *sub-differential* $\partial h(\mathbf{x})$
  - $\mathbf{q} \in \partial h(\mathbf{x}) \iff h(\mathbf{y}) \geq h(\mathbf{x}) + \mathbf{q}^\top (\mathbf{y} - \mathbf{x})$.

Proximal GD converges at order $\mathcal{O}(\frac{1}{k})$

Theorem. For $\min_{\mathbf{x}} F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$ where

- f is L-smooth[2] and convex
- g is convex (possibly non-differentiable),

the sequence $\{F(\mathbf{x}_k)\}_{k\in\mathbb{N}}$ produced by proximal GD iteration

$$\mathbf{x}_{k+1} = \mathrm{prox}_{tg}\big(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)\big) = \mathbf{x}_k - t_k G_{t_k}(\mathbf{x}_k)$$

converges to $F^* = F(\mathbf{x}^*)$ at the rate $\mathcal{O}(\frac{1}{k})$ as

$$F(\mathbf{x}_k) - F^* \leq \frac{L\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2k}.$$

[2] ∇f is L-Lipschitz.

The proof ... (1/6)

- Consider f

$$f \text{ is } L\text{-smooth} \qquad f(\mathbf{y}) \le f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \tag{1}$$

$$f \text{ is convex} \qquad f(\mathbf{x}) \le f(\mathbf{z}) - \nabla f(\mathbf{x})^\top (\mathbf{z}-\mathbf{x}) \tag{2}$$

$$(1)+(2) \qquad f(\mathbf{y}) \le f(\mathbf{z}) - \nabla f(\mathbf{x})^\top (\mathbf{z}-\mathbf{y}) + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 \tag{3}$$

- Put the gradient mapping $\mathbf{y} = \mathbf{x} - tG_t(\mathbf{x})$ into (3)

$$f\big(\mathbf{x}-tG_t(\mathbf{x})\big) \;\le\; f(\mathbf{z}) - \nabla f(\mathbf{x})^\top \big(\mathbf{z}-\mathbf{x}+tG_t(\mathbf{x})\big) + \frac{Lt^2}{2}\|G_t(\mathbf{x})\|_2^2 \tag{4}$$

- As $F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$, by (4)

$$F\big(\mathbf{x}-tG_t(\mathbf{x})\big) \le f(\mathbf{z}) - \nabla f(\mathbf{x})^\top \big(\mathbf{z}-\mathbf{x}+tG_t(\mathbf{x})\big) + \frac{Lt^2}{2}\|G_t(\mathbf{x})\|_2^2 + g\big(\mathbf{x}-tG_t(\mathbf{x})\big)$$

- What's next: handle the term $g\big(\mathbf{x}-tG_t(\mathbf{x})\big)$

The proof ... (2/6)

- To handle $g\big(\mathbf{x} - tG_t(\mathbf{x})\big)$, we show

$$g\big(\mathbf{x} - tG_t(\mathbf{x})\big) \leq g(\mathbf{z}) - (G_t(\mathbf{x}) - \nabla f(\mathbf{x}))^\top \Big(\mathbf{z} - \mathbf{x} + tG_t(\mathbf{x})\Big). \qquad (\#)$$

- The key is subgradient: g is non-differentiable but convex, so subgradient inequality holds

$$g(\mathbf{z}) \geq g(\mathbf{x} - tG_t(\mathbf{x})) + \mathbf{q}^\top \Big(\mathbf{z} - (\mathbf{x} - tG_t(\mathbf{x}))\Big) \qquad (*)$$

  i.e., there is an affine global underestimator formed by $\mathbf{q}$ (the subgradient) at point $\mathbf{x} - tG_t(\mathbf{x})$ for all $\mathbf{z}$.

- Re-arrange (*) gives

$$g(\mathbf{x} - tG_t(\mathbf{x})) \leq g(\mathbf{z}) - \mathbf{q}^\top \Big(\mathbf{z} - \mathbf{x} + tG_t(\mathbf{x})\Big). \qquad (\#\#)$$

- Now the question is : how to find the subgradient $\mathbf{q}$?
  Answer: Comparing $(\#)$ and $(\#\#)$ we see that we need to show $G_t(\mathbf{x}) - \nabla f(\mathbf{x})$ is a subgradient of g. To do so we need to use the optimality condition.

$$g(\mathbf{x} - tG_t(\mathbf{x})) \leq g(\mathbf{z}) - (G_t(\mathbf{x}) - \nabla f(\mathbf{x}))^\top \Big(\mathbf{z} - \mathbf{x} + tG_t(\mathbf{x})\Big)$$

- Recall the equivalence between gradient mapping and the proximal operator

$$\mathbf{w}^* = \mathbf{x} - tG_t(\mathbf{x}) = \operatorname{prox}_{tg}\Big(\mathbf{x} - t\nabla f(\mathbf{x})\Big) = \arg\min_{\mathbf{w}} g(\mathbf{w}) + \frac{1}{2t}\|\mathbf{w} - \mathbf{x} + t\nabla f(\mathbf{x})\|_2^2 \qquad (**)$$

- The subgradient optimality condition of (**) is $0 \in \partial g(\mathbf{w}^*) + \partial \frac{1}{2t}\|\mathbf{w}^* - \mathbf{x} + t\nabla f(\mathbf{x})\|_2^2$. Once again, why subgradient is used here is because g is (possibly) not differentiable. Note that $\|\mathbf{w}^* - \mathbf{x} + t\nabla f(\mathbf{x})\|_2^2$ is differentiable, its subgradient can be replaced by the normal gradient, so now the subgradient optimality condition of (**) becomes

$$0 \in \partial g(\mathbf{w}^*) + \frac{1}{t}(\mathbf{w}^* - \mathbf{x} + t\nabla f(\mathbf{x})) \quad \Longleftrightarrow \quad \frac{1}{t}(\mathbf{x} - \mathbf{w}^*) - \nabla f(\mathbf{x}) \in \partial g(\mathbf{w}^*).$$

- $\frac{1}{t}(\mathbf{x} - \mathbf{w}^*) \overset{(**)}{=} G_t(\mathbf{x})$ hence $G_t(\mathbf{x}) - \nabla f(\mathbf{x}) \in \partial g(\mathbf{w}^*)$, i.e., $G_t(\mathbf{x}) - \nabla f(\mathbf{x})$ is a subgradient of g at $\mathbf{w}^*$, therefore

$$g(\mathbf{x} - tG_t(\mathbf{x})) \leq g(\mathbf{z}) - (G_t(\mathbf{x}) - \nabla f(\mathbf{x}))^\top \Big(\mathbf{z} - \mathbf{x} + tG_t(\mathbf{x})\Big).$$

The proof ... (3/6)

We now have

$$\begin{aligned} F\big(\mathbf{x} - tG_t(\mathbf{x})\big) &\leq f(\mathbf{z}) - \nabla f(\mathbf{x})^\top \big(\mathbf{z} - \mathbf{x} + tG_t(\mathbf{x})\big) + \frac{Lt^2}{2}\|G_t(\mathbf{x})\|_2^2 + g\big(\mathbf{x} - tG_t(\mathbf{x})\big) \\ g(\mathbf{x} - tG_t(\mathbf{x})) &\leq g(\mathbf{z}) - (G_t(\mathbf{x}) - \nabla f(\mathbf{x}))^\top \Big(\mathbf{z} - \mathbf{x} + tG_t(\mathbf{x})\big)\Big) \end{aligned}$$

Add the two

$$\begin{aligned} F\Big(\mathbf{x} - tG_t(\mathbf{x})\Big) &\leq f(\mathbf{z}) + g(\mathbf{z}) - G_t(\mathbf{x})^\top \Big(\mathbf{z} - \mathbf{x} + tG_t(\mathbf{x})\Big) + \frac{Lt^2}{2}\|G_t(\mathbf{x})\|_2^2 \\ \begin{pmatrix} F = f + g, \\ \text{put } \mathbf{z} = \mathbf{x}^* \end{pmatrix} &= F^* - G_t(\mathbf{x})^\top \Big(\mathbf{x}^* - \mathbf{x} + tG_t(\mathbf{x})\Big) + \frac{Lt^2}{2}\|G_t(\mathbf{x})\|_2^2 \\ (\text{let } t \leq \frac{1}{L}) &\leq F^* - G_t(\mathbf{x})^\top \Big(\mathbf{x}^* - \mathbf{x} + tG_t(\mathbf{x})\Big) + \frac{t}{2}\|G_t(\mathbf{x})\|_2^2 \\ (\text{combine } G^\top G) &= F^* - G_t(\mathbf{x})^\top (\mathbf{x}^* - \mathbf{x}) - \frac{t}{2}\|G_t(\mathbf{x})\|_2^2 \end{aligned}$$

The proof ... (4/6)

- The most tricky step, pure algebra trick

$$-G_t(\mathbf{x})^\top(\mathbf{x}^* - \mathbf{x}) - \frac{t}{2}\|G_t(\mathbf{x})\|_2^2 = \frac{1}{2t}\Big(\|\mathbf{x} - \mathbf{x}^*\|_2^2 - \|\mathbf{x} - \mathbf{x}^* - tG_t(\mathbf{x})\|_2^2\Big)$$

(note that $\mathbf{x}^* - \mathbf{x}$ becomes $\mathbf{x} - \mathbf{x}^*$ in the second term on right hand side)

- Hence

$$\begin{aligned} F\Big(\mathbf{x} - tG_t(\mathbf{x})\Big) &\leq F^* - G_t(\mathbf{x})^\top(\mathbf{x}^* - \mathbf{x}) - \frac{t}{2}\|G_t(\mathbf{x})\|_2^2 \\ &\leq F^* + \frac{1}{2t}\Big(\|\mathbf{x} - \mathbf{x}^*\|_2^2 - \|\mathbf{x} - \mathbf{x}^* - tG_t(\mathbf{x})\|_2^2\Big) \end{aligned}$$

- Put $\mathbf{x} = \mathbf{x}_k$ and recall $\mathbf{x}_k - t_k G_{t_k}(\mathbf{x}_k) = \mathbf{x}_{k+1}$

$$F(\mathbf{x}_{k+1}) - F^* \leq \frac{1}{2t_k}\Big(\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2\Big).$$

It forms a telescoping series !

The proof ... (5/6)

- Compute for each k

$$\begin{aligned}
k = 0 \qquad & F(\mathbf{x}_1) - F^* \leq \frac{1}{2t_0}\Big(\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_1 - \mathbf{x}^*\|_2^2\Big) \\
k = 1 \qquad & F(\mathbf{x}_2) - F^* \leq \frac{1}{2t_1}\Big(\|\mathbf{x}_1 - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_2 - \mathbf{x}^*\|_2^2\Big) \\
\vdots \qquad & \qquad\qquad\qquad\qquad \vdots \\
k = K-1 \qquad & F(\mathbf{x}_K) - F^* \leq \frac{1}{2t_{K-1}}\Big(\|\mathbf{x}_{K-1} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_K - \mathbf{x}^*\|_2^2\Big)
\end{aligned}$$

- Use constant stepsize $t_k = t \leq \dfrac{1}{L}$, sum all the inequalities gives

$$\sum_{i=0}^{K-1}\Big(F(\mathbf{x}_i) - F^*\Big) \leq \frac{1}{2t}\Big(\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_K - \mathbf{x}^*\|_2^2\Big) \leq \frac{1}{2t}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2.$$

The proof ... (6/6)

- By the decent lemma of proximal gradient (see p.8 here), we have a non-increasing sequence of F: $F(\mathbf{x}_K) \leq ... \leq F(\mathbf{x}_i) \leq ... \leq F(\mathbf{x}_2) \leq F(\mathbf{x}_1) \leq F(\mathbf{x}_0)$, so

$$\sum_{i=0}^{K-1} \Big(F(\mathbf{x}_K) - F^* \Big) \leq \sum_{i=0}^{K-1} \Big(F(\mathbf{x}_i) - F^* \Big) \leq \frac{1}{2t} \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2$$

- Ignore the middle term

$$\sum_{i=0}^{K-1} \Big(F(\mathbf{x}_K) - F^* \Big) \leq \frac{1}{2t} \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2$$

- As $F(\mathbf{x}_K)$ and F^* are independent of the subscript i,

$$K \big(F(\mathbf{x}_K) - F^* \big) \leq \frac{1}{2t} \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2$$

- Re-arrange and put $t = \dfrac{1}{L}$

$$F(\mathbf{x}_K) - F^* \leq \frac{L \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2K}. \quad \square$$

Last page - summary

- Proximal operator of a function f at point $\mathbf{x}_0$:

$$\mathrm{prox}_f(\mathbf{x}_0) = \arg\min_{\mathbf{x}} \left\{ f(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2 \right\}.$$

- Proximal GD solves $\min F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$ where f, g convex but f differentiable g is not.
- Proximal GD update $\mathbf{x}_{k+1} = \mathrm{prox}_{t_k g}(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k))$
- Gradient mapping $G_t(\mathbf{x}) = \frac{1}{t}\Big[\mathbf{x} - \mathrm{prox}_{tg}(\mathbf{x} - t\nabla f(\mathbf{x}))\Big]$
- Proximal GD with fix stepsize $t_k = \frac{1}{L}$ converges as $F(\mathbf{x}_k) - F^* \leq \frac{L\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2k}$.

End of document