CS 2601 Linear and Convex Optimization

8. Proximal gradient descent

Bo Jiang

John Hopcroft Center for Computer Science
Shanghai Jiao Tong University

Fall 2022

Outline

- Algorithm and examples
- Convergence analysis

Composite functions

Consider

$$F(\boldsymbol{x}) = f(\boldsymbol{x}) + h(\boldsymbol{x})$$

where

- f is convex and smooth
- h is convex but not necessarily smooth

Example. Model with ℓ_1 regularization to promote sparsity,

$$F(\boldsymbol{x}) = f(\boldsymbol{x}) + \lambda\|\boldsymbol{x}\|_1$$

A more concrete example is Lasso in penalized form,

$$F(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_1$$

Proximal gradient descent

Recall for differentiable F, the update rule for gradient descent is

$$\boldsymbol{x}_{k+1} = \operatorname*{argmin}_{\boldsymbol{x}} \left\{ \underbrace{F(\boldsymbol{x}_k) + \nabla F(\boldsymbol{x}_k)^T(\boldsymbol{x} - \boldsymbol{x}_k) + \frac{1}{2t_k}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2}_{\hat{F}(\boldsymbol{x})} \right\}$$

加上常数凑平方

$$= \operatorname*{argmin}_{\boldsymbol{x}} \ \frac{1}{2t_k}\|\boldsymbol{x} - (\boldsymbol{x}_k - t_k \nabla F(\boldsymbol{x}_k))\|_2^2$$

Proximal gradient descent uses a similar approximation, but only for the smooth part f,

$$\boldsymbol{x}_{k+1} = \operatorname*{argmin}_{\boldsymbol{x}} \left\{ \underbrace{f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^T(\boldsymbol{x} - \boldsymbol{x}_k) + \frac{1}{2t_k}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2}_{\hat{f}(\boldsymbol{x})} + h(\boldsymbol{x}) \right\}$$

$$= \operatorname*{argmin}_{\boldsymbol{x}} \left\{ \frac{1}{2}\|\boldsymbol{x} - (\boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k))\|_2^2 + t_k h(\boldsymbol{x}) \right\}$$

近邻梯度下降

"Proximal" means we try to make $\boldsymbol{x}_{k+1}$ stay close to $\boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k)$.

Proximal gradient descent (cont'd)

Define the proximal mapping or proximal operator,

$$\mathrm{prox}_h(\boldsymbol{y}) = \underset{\boldsymbol{x}}{\mathrm{argmin}} \left\{ \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 + h(\boldsymbol{x}) \right\}$$

so

$$\boldsymbol{x}_{k+1} = \mathrm{prox}_{t_k h}(\boldsymbol{x}_k - t_k \nabla f(\boldsymbol{x}_k))$$

Proximal gradient descent with constant step size

1: initialization $\boldsymbol{x} \leftarrow \boldsymbol{x}_0 \in \mathbb{R}^n$
2: **while** stopping criterion not satisfied **do**
3: $\quad \boldsymbol{x} \leftarrow \mathrm{prox}_{th}(\boldsymbol{x} - t\nabla f(\boldsymbol{x}))$
4: **end while**
5: **return** $\boldsymbol{x}$

Note. The proximal operator involves another optimization problem! Fortunately, it depends only on h not on f, and we can compute it in closed form for many important h, e.g. ℓ_1 regularization.

Proximal operator for ℓ_2 regularization

When $h(\boldsymbol{x}) = \frac{\lambda}{2}\|\boldsymbol{x}\|_2^2$, the proximal operator is

$$\mathrm{prox}_h(\boldsymbol{y}) = \operatorname*{argmin}_{\boldsymbol{x}} \left\{ \overset{\text{fixed part}}{\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2} + \overset{\text{variant}}{\frac{\lambda}{2}\|\boldsymbol{x}\|_2^2} \right\} = \frac{\boldsymbol{y}}{1+\lambda}$$

Note. ℓ_2 regularization does not promote sparsity.

For ℓ_2 regularization, proximal gradient descent turns out to be nothing but gradient descent.

The proximal gradient step with step size t is

$$\boldsymbol{x}_{k+1} = \frac{\boldsymbol{x}_k - t\nabla f(\boldsymbol{x}_k)}{1+\lambda t} = \boldsymbol{x}_k - \frac{t}{1+\lambda t}(\nabla f(\boldsymbol{x}_k) + \lambda \boldsymbol{x}_k) = \boldsymbol{x}_k - \frac{t}{1+\lambda t}\nabla F(\boldsymbol{x}_k)$$

exactly the gradient step for $F(\boldsymbol{x}) = f(\boldsymbol{x}) + h(\boldsymbol{x})$ with step size $\frac{t}{1+\lambda t}$!

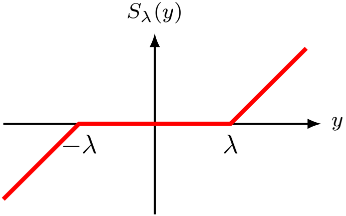
Proximal operator for ℓ_1 regularization

First consider the 1D case, $h(x) = \lambda|x|$ with $\lambda \geq 0$. Need to solve

$$\min_x \quad \frac{1}{2}(x-y)^2 + \lambda|x|$$

The proximal operator is given by the soft-thresholding operator $S_\lambda(y)$,

$$\operatorname{prox}_{\lambda|\cdot|}(y) = S_\lambda(y) = \operatorname{sgn}(y)(|y|-\lambda)^+ = \begin{cases} y-\lambda, & \text{if } y > \lambda \\ 0, & \text{if } -\lambda \leq y \leq \lambda \\ y+\lambda, & \text{if } y < -\lambda \end{cases}$$



Proximal operator for ℓ_1 regularization (cont'd)

Proof. Let $x^* = \mathrm{prox}_h(y)$ to be the minimum of $\frac{1}{2}(x-y)^2 + \lambda|x|$. Then

$$\frac{1}{2}(x^*-y)^2 + \lambda|x^*| \leq \frac{1}{2}(-x^*-y)^2 + \lambda|-x^*| \implies yx^* \geq 0$$

i.e. x^* must have the same sign as y. If $y \geq 0$, then $x^* \geq 0$, so

$$x^* = \operatorname*{argmin}_{x \geq 0} \left\{ \frac{1}{2}(x-y)^2 + \lambda x \right\} = \begin{cases} y-\lambda, & \text{if } y > \lambda \\ 0, & \text{if } 0 \leq y \leq \lambda \end{cases}$$

Similarly, if $y \leq 0$, then $x^* \leq 0$, so

$$x^* = \operatorname*{argmin}_{x \leq 0} \left\{ \frac{1}{2}(x-y)^2 - \lambda x \right\} = \begin{cases} y+\lambda, & \text{if } y < -\lambda \\ 0, & \text{if } -\lambda \leq y \leq 0 \end{cases}$$

Combining the two cases completes the proof.

Proximal operator for ℓ_1 regularization (cont'd)

When $h(\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|_1$, the minimization in the proximal operator can be decomposed into n 1D problems, one for each component,

$$\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n}\left\{\frac{1}{2}(x_i - y_i)^2 + \lambda|x_i|\right\}$$

The proximal operator simply applies the soft-thresholding operator to each component of $\boldsymbol{y}$,

$$\left[\operatorname{prox}_{\lambda\|\cdot\|_1}(\boldsymbol{y})\right]_i = [S_\lambda(\boldsymbol{y})]_i = S_\lambda(y_i) = \begin{cases} y_i - \lambda, & \text{if } y_i > \lambda \\ 0, & \text{if } |y_i| \leq \lambda \\ y_i + \lambda, & \text{if } y_i < -\lambda \end{cases}$$

Note. The soft-thresholding operation gives us some intuition about how ℓ_1 regularization promotes sparsity: small components are set to zero during the optimization process.

Lasso and ISTA

For Lasso in the penalized form,

$$\min_{\boldsymbol{w}} \quad F(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{w}\|_1$$

the smooth part $f(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$ has gradient

$$\nabla f(\boldsymbol{w}) = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})$$

so the proximal gradient step is

$$\boldsymbol{w}_{k+1} = S_{\lambda t}(\boldsymbol{w}_k - t\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w}_k - \boldsymbol{y}))$$

Known as the iterative soft-thresholding algorithm (ISTA).

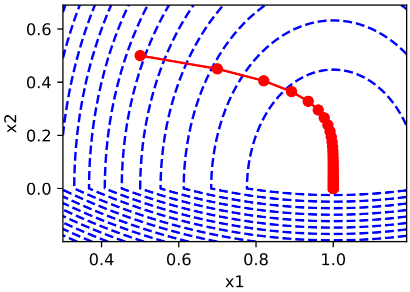
Note. The threshold is λt, not λ.

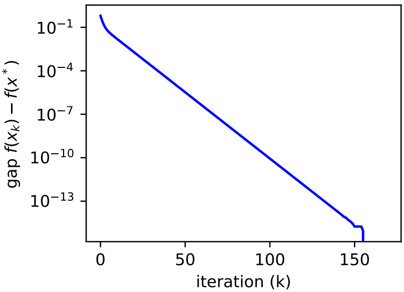
Lasso and ISTA (cont'd)

Example.

$$X = \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad y = \begin{bmatrix} 3 \\ 2 \\ 2 \end{bmatrix}, \quad \lambda = 2$$

Step size $t = 0.1$, $w_0 = (0.5, 0.5)^T$, $w^* = (1, 0)^T$.





Outline

- Algorithm and examples
- Convergence analysis

Lower bound for strongly convex functions

Lemma. If $\varphi : X \to \mathbb{R}$ is μ-strongly convex with a minimum $\boldsymbol{x}^*$, then

$$\varphi(\boldsymbol{x}) \geq \varphi(\boldsymbol{x}^*) + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2, \quad \forall \boldsymbol{x} \in X$$

Note. For φ is differentiable, this will follow from the quadratic lower bound $\varphi(\boldsymbol{x}) \geq \varphi(\boldsymbol{y}) + \nabla\varphi(\boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y}) + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$ and the optimality condition $\nabla\varphi(\boldsymbol{x}^*)^T(\boldsymbol{x} - \boldsymbol{x}^*) \geq 0$.

如果连续可以直接用upperbound证明，但是不需要

Proof. Since φ is μ-strongly convex, there exists a convex $\tilde{\varphi}(\boldsymbol{x})$ s.t.

$$\varphi(\boldsymbol{x}) = \tilde{\varphi}(\boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|^2$$

后项展开，不会影响convexity

Fix $\boldsymbol{x}$ and let $\boldsymbol{x}_t = t\boldsymbol{x} + \bar{t}\boldsymbol{x}^* = \boldsymbol{x}^* + t(\boldsymbol{x} - \boldsymbol{x}^*)$, $t \in [0, 1]$. By convexity of $\tilde{\varphi}$,

$$\varphi(\boldsymbol{x}^*) \leq \varphi(\boldsymbol{x}_t) = \tilde{\varphi}(\boldsymbol{x}_t) + \frac{\mu}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^*\|_2^2 \leq t\tilde{\varphi}(\boldsymbol{x}) + \bar{t}\tilde{\varphi}(\boldsymbol{x}^*) + \frac{\mu t^2}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2$$

Since $\tilde{\varphi}(\boldsymbol{x}^*) = \varphi(\boldsymbol{x}^*)$,

$$\tilde{\varphi}(\boldsymbol{x}) \geq \varphi(\boldsymbol{x}^*) - \frac{\mu t}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2, \ \forall t \in [0, 1] \implies \tilde{\varphi}(\boldsymbol{x}) \geq \varphi(\boldsymbol{x}^*)$$

Convergence analysis

Theorem. Let $F(\boldsymbol{x}) = f(\boldsymbol{x}) + h(\boldsymbol{x})$, where h is convex, f is L-smooth and m-strongly convex with $m \geq 0$. Let $\boldsymbol{x}^*$ be the minimum of F. The sequence $\{\boldsymbol{x}_k\}$ produced by proximal gradient descent with constant step size $t = \frac{1}{L}$ has the following properties.

1. $F(\boldsymbol{x}_{k+1}) \leq F(\boldsymbol{x}_k)$ and

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

2. $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|_2^2 \leq (1 - \frac{m}{L}) \|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2$, and hence

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 \leq (1 - \frac{m}{L})^k \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2$$

Note. If $t \in (0, \frac{1}{L}]$, then the conclusions hold with L replaced by $\frac{1}{t}$, since an L-smooth function is also $\frac{1}{t}$-smooth.

Note. 2 implies $F(\boldsymbol{x}_k) - F(\boldsymbol{x}^*) \leq C(1 - \frac{m}{L})^{\frac{k}{2}} \|\boldsymbol{x}^* - \boldsymbol{x}_0\|_2$ for C-Lipschitz continuous F.

Proof

skipped

Let

$$\widehat{F}(\boldsymbol{x}) = f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^T(\boldsymbol{x} - \boldsymbol{x}_k) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2 + h(\boldsymbol{x})$$

Since f is L-smooth and m-strongly convex,

$$\frac{m}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2 \leq f(\boldsymbol{x}) - f(\boldsymbol{x}_k) - \nabla f(\boldsymbol{x}_k)^T(\boldsymbol{x} - \boldsymbol{x}_k) \leq \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2$$

Plugging into $\widehat{F}(\boldsymbol{x})$ yields

$$F(\boldsymbol{x}) \leq \widehat{F}(\boldsymbol{x}) \leq F(\boldsymbol{x}) + \frac{L-m}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2$$

Note $\widehat{F}(\boldsymbol{x})$ is L-strongly convex, and $\boldsymbol{x}_{k+1} = \operatorname{argmin} \widehat{F}(\boldsymbol{x})$ when $t = \frac{1}{L}$.

$$\begin{aligned} F(\boldsymbol{x}_{k+1}) \leq \widehat{F}(\boldsymbol{x}_{k+1}) &\overset{(\star)}{\leq} \widehat{F}(\boldsymbol{x}) - \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{x}_{k+1}\|_2^2 \\ &\leq F(\boldsymbol{x}) + \frac{L-m}{2}\|\boldsymbol{x} - \boldsymbol{x}_k\|_2^2 - \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{x}_{k+1}\|_2^2 \end{aligned}$$

where $(\star)$ uses the previous lemma.

Proof (cont'd)

The previous slide shows

$$F(\boldsymbol{x}_{k+1}) \le F(\boldsymbol{x}) + \frac{L-m}{2}\|\boldsymbol{x}-\boldsymbol{x}_k\|_2^2 - \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{x}_{k+1}\|_2^2$$

1. Setting $\boldsymbol{x} = \boldsymbol{x}_k$ shows

$$F(\boldsymbol{x}_{k+1}) \le F(\boldsymbol{x}_k) - \frac{L}{2}\|\boldsymbol{x}_k-\boldsymbol{x}_{k+1}\|_2^2 \le F(\boldsymbol{x}_k)$$

Since $m \ge 0$, setting $\boldsymbol{x} = \boldsymbol{x}^*$ shows

$$F(\boldsymbol{x}_{i+1}) - F(\boldsymbol{x}^*) \le \frac{L}{2}\|\boldsymbol{x}^*-\boldsymbol{x}_i\|_2^2 - \frac{L}{2}\|\boldsymbol{x}^*-\boldsymbol{x}_{i+1}\|_2^2$$

Since $F(\boldsymbol{x}_{i+1}) \le F(\boldsymbol{x}_i)$ by part 1, summing over i from 0 to $k-1$,

$$\begin{aligned} F(\boldsymbol{x}_k) - F(\boldsymbol{x}^*) &\le \frac{1}{k}\sum_{i=0}^{k-1}[F(\boldsymbol{x}_i) - F(\boldsymbol{x}^*)] \\ &\le \frac{L}{2k}\|\boldsymbol{x}^*-\boldsymbol{x}_0\|_2^2 - \frac{L}{2k}\|\boldsymbol{x}^*-\boldsymbol{x}_k\|_2^2 \le \frac{L}{2k}\|\boldsymbol{x}^*-\boldsymbol{x}_0\|_2^2 \end{aligned}$$

Proof (cont'd)

2. Setting $\boldsymbol{x} = \boldsymbol{x}^*$ shows

$$\begin{aligned}
\|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\|_2^2 &\le (1 - \frac{m}{L})\|\boldsymbol{x}^* - \boldsymbol{x}_k\|_2^2 - \frac{2}{L}[\underbrace{F(\boldsymbol{x}_{k+1}) - F(\boldsymbol{x}^*)}_{\ge 0}] \\
&\le (1 - \frac{m}{L})\|\boldsymbol{x}^* - \boldsymbol{x}_k\|_2^2
\end{aligned}$$

so

$$\|\boldsymbol{x}^* - \boldsymbol{x}_k\|_2^2 \le (1 - \frac{m}{L})^k \|\boldsymbol{x}^* - \boldsymbol{x}_0\|_2^2$$