Proximal Gradient methods(continued)

Chenglong Bao

YMSC, Tsinghua University

Acknowledgement: slides based on Prof. Lieven Vandenberghes and Prof. Zaiwen Wen (PKU).

Inertial proximal algorithm

Consider the problem:

$$\min \quad f(x) = g(x) + h(x)$$

where ∇g is L-Lipschitz.

- Choose x_0 and set $x_{-1} = x^0$, choose $\beta \in [0,1]$, set $\alpha < 2(1-\beta)/L$ and computes $\quad x_{k+1} = \text{prox}_{\alpha h}(x_k - \alpha \nabla g(x_k))$

$y_k = x_k + \beta(x_k - x_{k-1})$.

$$x_{k+1} = \mathbf{prox}_{\alpha h}(x_k - \alpha \nabla g(x_k) + \underbrace{\beta(x_k - x_{k-1})})$$

- The term $\beta(x_k - x_{k-1})$: inertial term $\quad x_{k+1} = \text{prox}_{\alpha h}(y_k - \alpha \nabla g(x_k))$
- For $h = 0$, the scheme is referred as the Heavy ball method.
- Ref: P. Ochs, Y. Chen, T. Brox and T. Pock. IPiano: Inertial proximal algorithm for nonconvex optimization, SIAM J. Imaging Sciences, Vol 7, No.2.

Conditional gradient method: Motivation

Let $\mathcal{X}$ be a compact set and consider

$$\min_{x\in\mathcal{X}} \quad f(x)$$

- Proximal gradient method:

$$x^{k+1} = \arg\min_{x\in\mathcal{X}} \left\{ \overbrace{f(x_k) + \langle \nabla f(x_k), \overset{S}{\overset{\parallel}{\boxed{x - x_k}}} \rangle}^{\text{first order Approximation.}} + \overset{X.}{\boxed{\frac{1}{2\alpha_k}\|x - x_k\|^2}} \right\}$$

  It is equivalent to the projected gradient method:

$$x_{k+1} = \mathcal{P}_{\mathcal{X}}(x_k - \alpha_k \nabla f(x_k))$$

- Difficulty: computation of the projection $\mathcal{P}_{\mathcal{X}}(\cdot)$ may be expensive.

Conditional gradient (CndG) or Frank-Wolfe method

Given $y_0 = x_0$ and $\alpha_k \in (0, 1]$, the CndG methods takes

$$\boxed{x_k = \underset{x \in \mathcal{X}}{\arg\min} \langle \nabla f(y_{k-1}), x \rangle,} \quad y_k = (1 - \alpha_k) y_{k-1} + \alpha_k x_k$$

Supporting function on $\mathcal{X}$.

$$= y_{k-1} + \alpha_k (x_k - y_{k-1})$$

- diminishing step sizes:

$$\alpha_k = \frac{2}{k+1}$$

- Exact line search

$\nabla f(y_{k-1}) = z_{k-1}$

$$\alpha_k = \underset{\alpha \in [0,1]}{\arg\min} f((1 - \alpha) y_{k-1} + \alpha x_k)$$

$\min_{x \in \mathcal{X}} \langle z_{k-1}, x \rangle$

$\Leftrightarrow \min_x \langle z_{k-1}, x \rangle + \delta_{\mathcal{X}}(x) = -\max_x \{ -\langle z_{k-1}, x \rangle - \delta_{\mathcal{X}}(x) \}$

$= -\delta^*_{\mathcal{X}}(-z_{k-1})$

Examples

考虑带某一范数$\|\cdot\|$约束的凸优化问题，

$$\min_x f(x) \quad \text{s.t.} \quad \|x\| \le t.$$

用条件梯度法求解该问题时，需要计算子问题，

$$\begin{aligned} x_k &\in \underset{\|x\|\le t}{\operatorname{argmin}} \langle \nabla f(y_{k-1}), x \rangle \\ &= -t \cdot \left(\underset{\|x\|\le 1}{\arg\max} \langle \nabla f(y_{k-1}), x \rangle \right) \\ &= -t \cdot \partial \|\nabla f(y_{k-1})\|_*. \end{aligned} \tag{4}$$

其中$\|z\|_* = \sup\{z^T x, \|x\| \le 1\}$是$\|\cdot\|$的对偶范数。注意到(4) 条件梯度法的子问题相当于计算一个对偶范数的次梯度。如果计算$\|\cdot\|$范数的次梯度比计算在约束集合$X = \{x \in \mathbb{R}^n : \|x\| \le t\}$上的投影要简单，条件梯度法比投影梯度法效率更高。

$$\partial\|x\| = \{u \mid \langle u, x\rangle = \|x\|,\ \|u\|_* \le 1\}$$

由于ℓ_1 范数的对偶范数是ℓ_∞ 范数，因此用条件梯度法求解该问题时子问题为，

$$x_k \in -t\cdot\partial\|\nabla f(y_{k-1})\|_\infty.$$

考虑到ℓ_∞ 范数的次梯度为$\partial\|x\|_\infty = \{v : \langle v, x\rangle = \|x\|_\infty, \|v\|_1 \le 1\}$，子问题等价于，

$$\begin{cases} i_k \in \underset{i=1,\dots,n}{\operatorname{argmax}} \ |\nabla_i f(y_{k-1})| \\ x_k = -t\cdot \mathrm{sgn}\,[\nabla_{i_k} f(y_{k-1})]\cdot e_{i_k}. \end{cases}$$

其中$\nabla_i f(y_{k-1})$ 表示向量$\nabla f(y_{k-1})$ 的第i 个元素，e_i 表示第i 个元素为1的单位向量。可以看到计算$\|\cdot\|_\infty$ 的次梯度和计算集合$X := \{x\in\mathbb{R}^n : \|x\|_1 \le t\}$ 上的投影都需要$\mathcal{O}(n)$ 的计算复杂度，但是条件梯度法子问题计算明显要更简单直接。

Examples: ℓ_p 范数约束问题, $1 \le p \le \infty$

由于ℓ_p 范数的对偶范数是ℓ_q 范数，其中$1/p + 1/q = 1$，因此用条件梯度法求解该问题时子问题为，

$$x_k \in -t \cdot \partial \|\nabla f(y_{k-1})\|_q.$$

注意到ℓ_q 范数的次梯度为$\partial\|x\|_q = \{v : \langle v, x\rangle = \|x\|_q, \|v\|_p \le 1\}$，子问题等价于，

$$x_k^{(i)} = -\beta \cdot \mathrm{sgn}\left[\nabla_i f(y_{k-1})\right] \cdot |\nabla_i f(y_{k-1})|^{p/q}.$$

其中β 是使得$\|x_k\|_q = t$ 的归一化常数。可以看到，除过$p = 1, 2, \infty$ 这些特殊情形，条件梯度法的子问题计算复杂度比直接计算点在集合$X = \{x \in \mathbb{R}^n : \|x\|_p \le t\}$ 上的投影要简单，后者投影计算需要单独解一个优化问题。

Example: 矩阵核范数约束优化问题

矩阵核范数$\|\cdot\|_*$的对偶范数是其谱范数$\|\cdot\|_2$:

$$\|X\|_* = \sum_{i=1}^{\min\{m,n\}} \sigma_i(X), \qquad \|X\|_2 = \max_{i=1,\ldots,\min\{m,n\}} \sigma_i(X).$$

因此条件梯度法的子问题为$X_k \in -t \cdot \partial\|\nabla f(Y_{k-1})\|_2$. 对矩阵范数的次梯度: $\partial\|X\| = \{Y : \langle Y, X\rangle = \|X\|, \|Y\|_* \leq 1\}$，设$u, v$分别是矩阵$\nabla f(Y_{k-1})$最大奇异值对应的左、右奇异向量，注意到，

$$\langle uv^T, \nabla f(Y_{k-1})\rangle = u^T \nabla f(Y_{k-1}) v = \sigma_{\max}(\nabla f(Y_{k-1})) = \|\nabla f(Y_{k-1})\|_2.$$

且$\|uv^T\|_* = 1$，因此矩阵$uv^T \in \partial\|\nabla f(Y_{k-1})\|_2$。则条件梯度法子问题等价于，

$$X_k \in -t \cdot uv^T. \tag{5}$$

可以看到，条件梯度法计算子问题时只需要计算矩阵最大的奇异值对应的左、右奇异向量。如果采用投影梯度法，其子问题是计算X到集合$\{X \in \mathbb{R}^{m\times n} : \|X\|_* \leq t\}$的投影，需要对矩阵做全奇异值分解，计算量比条件梯度法复杂很多。

Convergence: Lemma

令$\gamma_t \in (0,1]$，$t = 1,2,\dots$，构造序列

$$\Gamma_t = \begin{cases} 1 & t = 1 \\ (1-\gamma_t)\Gamma_{t-1} & t \geq 2 \end{cases}.$$

$$= \frac{\prod_{j=1}^{t}(1-\gamma_j)}{1-\gamma_1} \quad \checkmark.$$

如果序列$\{\Delta_t\}_{t\geq 0}$ 满足

$$\star \quad \Delta_t \leq (1-\gamma_t)\Delta_{t-1} + B_t \quad t = 1,2,\dots.$$

则对任意的k 我们对Δ_k 有估计 $\leq (1-\gamma_t)\big((1-\gamma_{t-1})\Delta_{t-2} + B_{t-1}\big) + B_t$

$f_t - f^*$

$$\boxed{\Delta_k \leq \Gamma_k(1-\gamma_1)\Delta_0 + \Gamma_k \sum_{t=1}^{k} \frac{B_t}{\Gamma_t}.}$$

Δ

$f_k - f^* \leq \dots(f_0 - f^*) + \dots$

Convergence

Let $f(x)$ is convex, $\nabla f(x)$ is L-Lipschitz, $D_X = \sup_{x,y\in X}\|x-y\|$. Then

$$f(y_k) - f(x^*) \;\le\; \frac{2L}{k(k+1)}\sum_{i=1}^{k}\|x_i - y_{i-1}\|^2 \;\le\; \frac{2L}{k+1}D_X^2. \quad O\left(\frac{1}{k}\right)$$

Proof: 令$\gamma_k = \frac{2}{k+1}$，记$\bar{y}_k = (1-\gamma_k)y_{k-1} + \gamma_k x_k$，则不管

$$\alpha_k = \frac{2}{k+1} \quad 或 \quad \alpha_k = \operatorname*{argmin}_{\alpha\in[0,1]} f((1-\alpha)y_{k-1} + \alpha x_k).$$

对$y_k = (1-\alpha_k)y_{k-1} + \alpha_k x_k$，我们都有$f(y_k) \le f(\bar{y}_k)$。注意到$\bar{y}_k - y_{k-1} = \gamma_k(x_k - y_{k-1})$，由$f(x) \in C_L^{1,1}(X)$有

$$f(y_k) \le f(\bar{y}_k) \;\le\; f(y_{k-1}) + \langle \nabla f(y_{k-1}), \bar{y}_k - y_{k-1}\rangle + \frac{L}{2}\|\bar{y}_k - y_{k-1}\|^2 \quad ✓ \tag{6}$$

$$\le \; (1-\gamma_k)[f(y_{k-1}) + \gamma_k[f(y_{k-1}) + \langle \nabla f(y_{k-1}), x - y_{k-1}\rangle] + \frac{L\gamma_k^2}{2}\|x_k - y_{k-1}\|^2 \tag{7}$$

$$\le \; (1-\gamma_k)f(y_{k-1}) + \gamma_k f(x) + \frac{L\gamma_k^2}{2}\|x_k - y_{k-1}\|^2, \quad 对任意\ x\in X. \tag{8}$$

Subproblem.

其中不等式(7) 是因为$x_k \in \min_{x\in X}\langle \nabla f(y_{k-1}), x\rangle$，由最优性条件我们可以得到对任意$x \in X$ 有$\langle x - x_k, \nabla f(y_{k-1})\rangle \geq 0$。将不等式(8) 稍做变换，对任意$x \in X$，

$$\underbrace{\overset{\Delta_k}{f(y_k) - f(x)}}_{\Delta} \leq (1-\gamma_k)[\overset{\Delta_{k-1}}{f(y_{k-1}) - f(x)}] + \frac{L}{2}\gamma_k^2 \overset{B_k}{\|x_k - y_{k-1}\|^2}. \qquad (9)$$

由引理可知，

$$f(y_k) - f(x) \leq \Gamma_k(1-\gamma_1)[f(y_0) - f(x)] + \frac{\Gamma_k L}{2}\sum_{i=1}^{k}\frac{\gamma_i^2}{\Gamma_i}\|x_i - y_{i-1}\|^2.$$

由$\gamma_k = \frac{2}{k+1}$，$\gamma_1 = 1$ 得到$\Gamma_k = \frac{2}{k(k+1)}$，我们可以得到收敛性不等式，

$$f(y_k) - f^* \leq \frac{2L}{\underset{\Delta}{k(k+1)}}\sum_{i=1}^{k}\|x_i - y_{i-1}\|^2 \leq \frac{2L}{k+1}D_X^2.$$

令$\frac{2L}{k+1}D_X^2 \leq \epsilon$，可以得到分析复杂度结论。

References

convergence analysis of proximal gradient method

- A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences (2009)
- A. Beck and M. Teboulle, *Gradient-based algorithms with applications to signal recovery*, in: Y. Eldar and D. Palomar (Eds.), *Convex Optimization in Signal Processing and Communications* (2009)