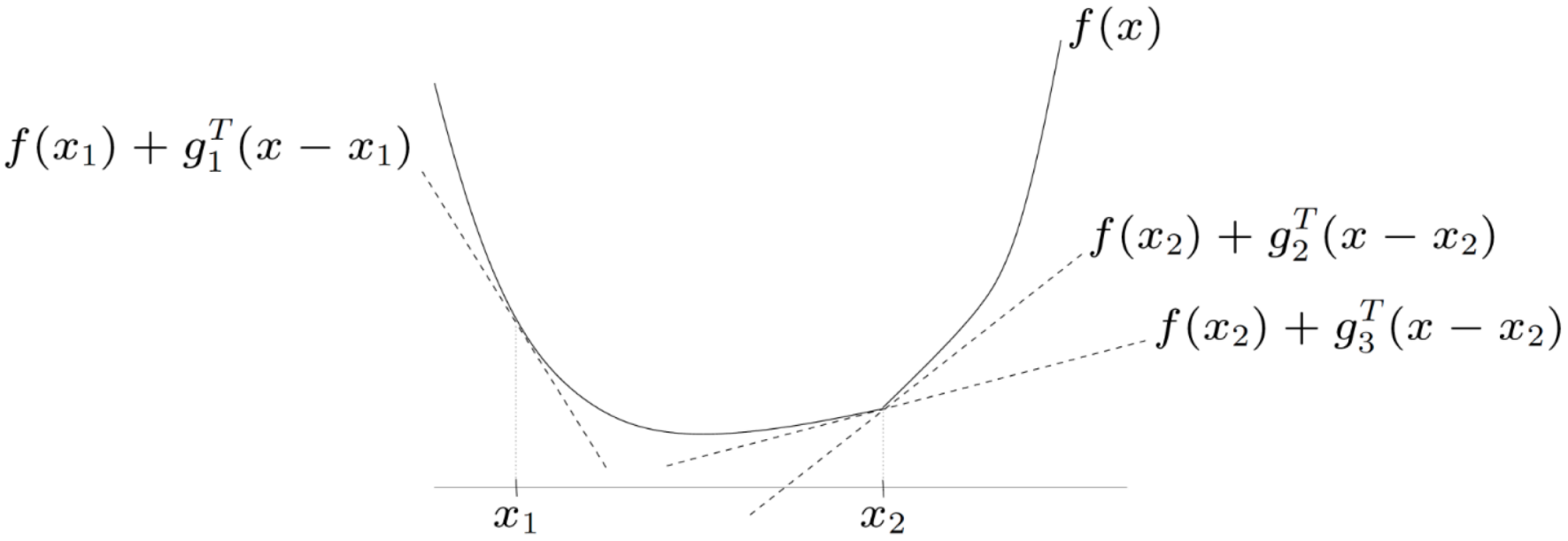
次梯度法(续)

王 尧

西安交通大学智能决策与机器学习中心
(Email: yao.s.wang@gmail.com)

2022. 5

回顾：次梯度与次微分



g_2, g_3 are subgradients at x_2; g_1 is a subgradient at x_1

the **subdifferential** $\partial f(x)$ of f at x is the set of all subgradients:

$$\partial f(x) = \{g | g^\top (y - x) \leq f(y) - f(x)\} \quad \forall\, y \in \mathbf{dom}\, f$$

回顾：对偶范数(dual norms)

- in the definition, $\|\cdot\|$ and $\|\cdot\|_*$ are a pair of dual norms:

$$\|u\|_* = \sup_{v \neq 0} \frac{u^T v}{\|v\|} = \sup_{\|v\|=1} u^T v$$

Examples:

1· The dual of the Euclidean norm is the Euclidean norm

2. The dual of the ℓ_∞ norm is the ℓ_1 norm (课后作业)

范数在零点的次梯度

Let $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$ for any norm $\|\cdot\|$, then for any $\boldsymbol{g}$ obeying $\|\boldsymbol{g}\|_* \leq 1$,

$$\boldsymbol{g} \in \partial f(\mathbf{0})$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$ (i.e. $\|\boldsymbol{x}\|_* := \sup_{\boldsymbol{z}:\|\boldsymbol{z}\|\leq 1} \langle \boldsymbol{z}, \boldsymbol{x} \rangle$)

Proof: To see this, it suffices to prove that

$$f(\boldsymbol{z}) \geq f(\mathbf{0}) + \langle \boldsymbol{g}, \boldsymbol{z} - \mathbf{0} \rangle, \qquad \forall \boldsymbol{z}$$

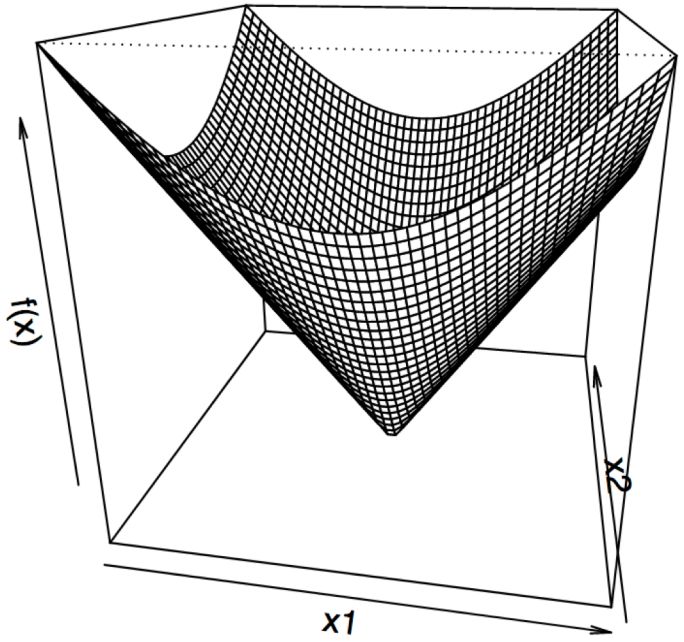
$$\iff \quad \langle \boldsymbol{g}, \boldsymbol{z} \rangle \leq \|\boldsymbol{z}\|, \qquad \forall \boldsymbol{z}$$

This follows from generalized Cauchy-Schwarz, i.e.

$$\langle \boldsymbol{g}, \boldsymbol{z} \rangle \leq \|\boldsymbol{g}\|_* \|\boldsymbol{z}\| \leq \|\boldsymbol{z}\|$$

L2范数的次梯度计算

Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_2$



- For $x \neq 0$, unique subgradient $g = x/\|x\|_2$
- For $x = 0$, subgradient g is any element of $\{z : \|z\|_2 \leq 1\}$

次梯度算法结构

为了极小化一个不可微的凸函数f，可类似梯度法构造如下次梯度算法的迭代格式：

$$x^{k+1} = x^k - \alpha_k g^k, \quad g^k \in \partial f(x^k),$$

其中$\alpha_k > 0$为步长.

1. 固定步长$\alpha_k = \alpha$；
2. 固定$\|x^{k+1} - x^k\|$，即$\alpha_k \|g^k\|$为常数
3. 消失步长$\alpha_k \to 0$且$\sum_{k=0}^{\infty} \alpha_k = +\infty$

基本假设

(1) f为凸函数；

(2) f至少存在一个有限的极小值点x^*，且$f(x^*) > -\infty$；

(3) f为利普希茨连续的，即

$$|f(x) - f(y)| \le G\|x - y\|, \quad \forall\, x, y \in \mathbb{R}^n,$$

其中$G > 0$为利普希茨常数.

我们下面证明这等价于$f(x)$的次梯度是有界的，即

$$\|g\| \le G, \quad \forall\, g \in \partial f(x), x \in \mathbb{R}^n.$$

证明：

- 充分性：假设$\|g\|_2 \leq G, \forall\, g \in \partial f(x)$；取$g_y \in \partial f(y), g_x \in \partial f(x)$:

$$g_x^{\mathrm{T}}(x-y) \geq f(x) - f(y) \geq g_y^{\mathrm{T}}(x-y)$$

再由柯西不等式

$$G\|x-y\|_2 \geq f(x) - f(y) \geq -G\|x-y\|_2$$

- 必要性：反设存在x和$g \in \partial f(x)$，使得$\|g\|_2 > G$；取$y = x + \frac{g}{\|g\|_2}$

$$\begin{aligned} f(y) &\geq f(x) + g^{\mathrm{T}}(y-x) \\ &= f(x) + \|g\|_2 \\ &> f(x) + G \end{aligned}$$

这与$f(x)$是G-利普希茨连续的矛盾.

基本不等式

- 次梯度方法不是一个下降方法，即无法保证$f(x^{k+1}) < f(x^k)$；
- 收敛性分析的关键是分析$f(x)$历史迭代的最优点所满足的性质.
- 设x^*是$f(x)$的一个全局极小值点，$f^* = f(x^*)$，根据迭代格式，

$$\begin{aligned}\left\|x^{i+1} - x^*\right\|^2 &= \left\|x^i - \alpha_i g^i - x^*\right\|^2 \\ &= \left\|x^i - x^*\right\|^2 - 2\alpha_i \left\langle g^i, x^i - x^*\right\rangle + \alpha_i^2 \left\|g^i\right\|^2 \\ &\leqslant \left\|x^i - x^*\right\|^2 - 2\alpha_i \left(f\left(x^i\right) - f^*\right) + \alpha_i^2 G^2\end{aligned}$$

- 结合$i = 0, \cdots, k$ 时相应的不等式，并定义$\hat{f}^k = \min_{0 \leqslant i \leqslant k} f\left(x^i\right)$：

$$\begin{aligned}2\left(\sum_{i=0}^{k} \alpha_i\right)\left(\hat{f}^k - f^*\right) &\leqslant \left\|x^0 - x^*\right\|^2 - \left\|x^{k+1} - x^*\right\|^2 + G^2 \sum_{i=0}^{k} \alpha_i^2 \\ &\leqslant \left\|x^0 - x^*\right\|^2 + G^2 \sum_{i=0}^{k} \alpha_i^2\end{aligned}$$

固定步长策略

(1) 取$\alpha_i = t$为固定步长，则

$$\hat{f}^k - f^* \leq \frac{\|x^0 - x^*\|^2}{2kt} + \frac{G^2 t}{2};$$

- $\hat{f}^k$ 无法保证收敛性
- 当k 足够大时,$\hat{f}^k$ 近似为$G^2 t/2$-次优的

(2) 取α_i使得$\|x^{i+1} - x^i\|$固定，即$\alpha_i \|g^i\| = s$为常数，则

$$\hat{f}^k - f^* \leq \frac{G\|x^0 - x^*\|^2}{2ks} + \frac{Gs}{2};$$

- $\hat{f}^k$ 无法保证收敛性
- 当k 足够大时,$\hat{f}^k$ 近似为$Gs/2$-次优的

消失步长策略

(3) 取α_i为消失步长，即$\alpha_i \to 0$且$\sum_{i=0}^{\infty} \alpha_i = +\infty$，则

$$\hat{f}^k - f^* \leq \frac{\|x^0 - x^*\|^2 + G^2 \sum_{i=0}^{k} \alpha_i^2}{2 \sum_{i=0}^{k} \alpha_i};$$

进一步可得$\hat{f}^k$收敛到f^*.

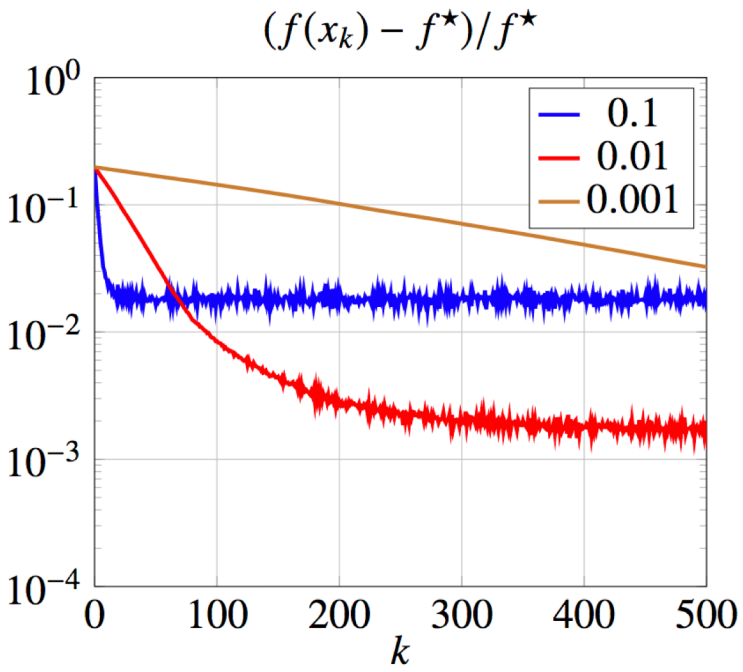
- 和梯度法不同，只有当α_k取消失步长时$\hat{f}^k$才具有收敛性.
- 一个常用的步长取法是$\alpha_k = \frac{1}{k}$.

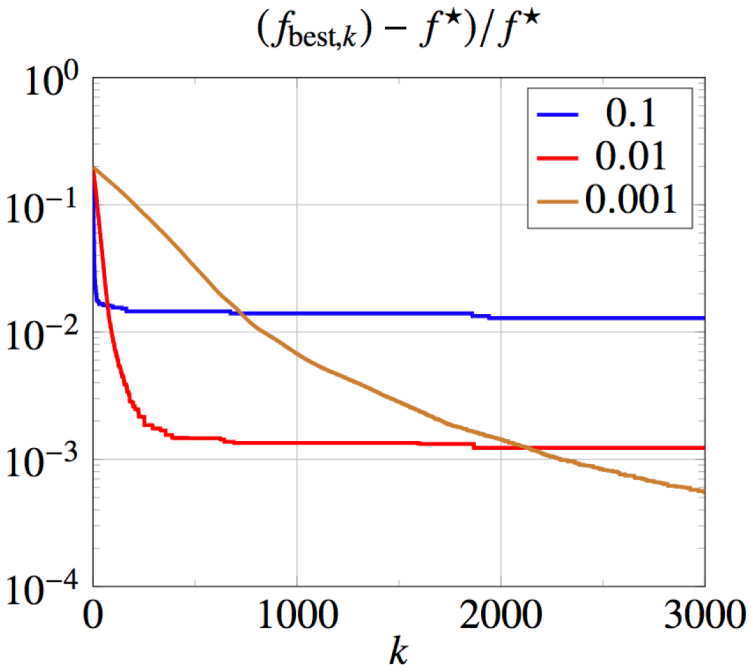
实例

$$\text{minimize } \|Ax - b\|_1$$

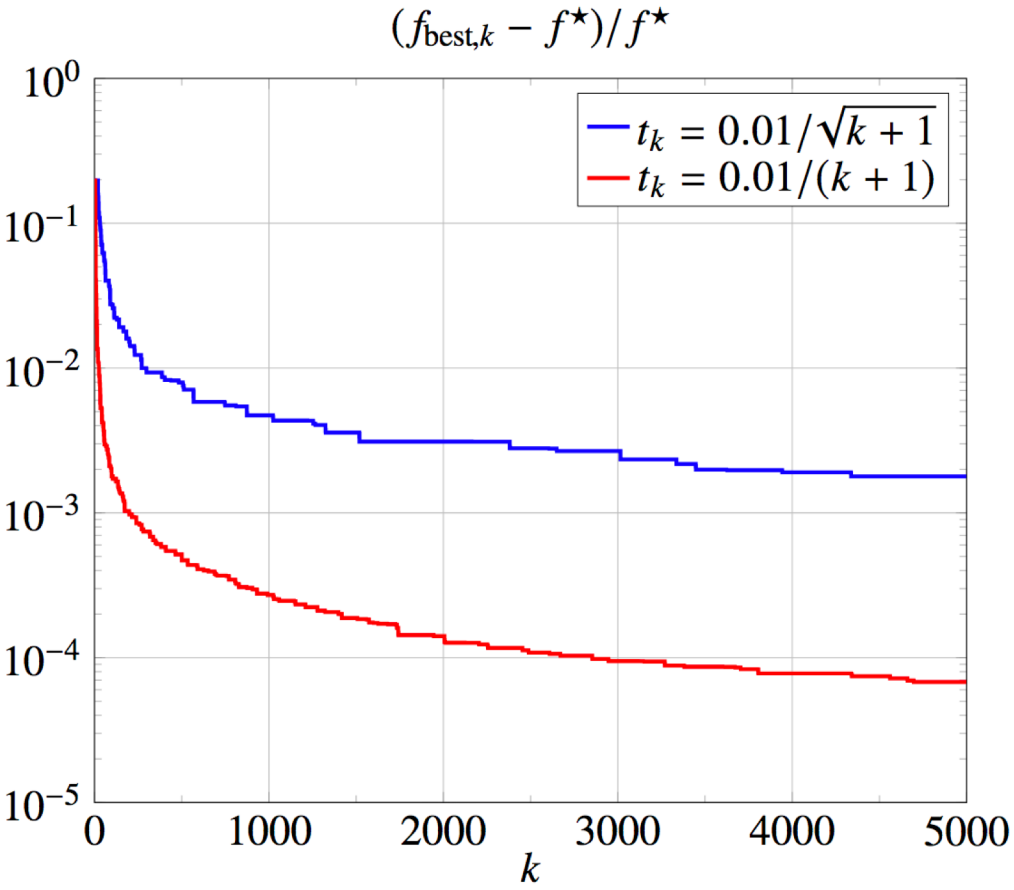
- example with $A \in \mathbf{R}^{500\times100}$, $b \in \mathbf{R}^{500}$

Fixed steplength $t_k = s/\|g_k\|_2$ for $s = 0.1, 0.01, 0.001$





Diminishing step size: $t_k = 0.01/\sqrt{k+1}$ and $t_k = 0.01/(k+1)$



收敛速率分析

- 假设$\left\|x^0 - x^*\right\| \leqslant R$，并且总迭代步数$k$是给定的，在固定步长下，

$$\hat{f}^k - f^* \leq \frac{\|x^0 - x^*\|^2}{2kt} + \frac{G^2 t}{2} \leq \frac{R^2}{2kt} + \frac{G^2 t}{2}.$$

- 由平均值不等式知当t满足$\frac{R^2}{2kt} = \frac{G^2 t}{2}$，即$t = \frac{R}{G\sqrt{k}}$时，右端达到最小.
- k步后得到的上界是

$$\hat{f}^k - f^* \leqslant \frac{GR}{\sqrt{k}}$$

- 这表明在$k = O(1/\epsilon^2)$步迭代后可以得到$\hat{f}^k - f^* \leq \epsilon$的精度

上述收敛速率远小于梯度法的收敛速率

思考：另外两种步长下的收敛速率如何？

总结

- 能够处理一般的不可微凸函数
- 常能推导出非常简单的算法
- 收敛速度可能非常缓慢
- 没有很好的停机准则
- 理论复杂度：迭代$O(1/\epsilon^2)$步，得到ϵ-次优的点

可分解函数

In words, we cannot do better than the $O(1/\epsilon^2)$ rate of subgradient method (unless we go beyond nonsmooth first-order methods)

So instead of trying to improve across the board, we will focus on minimizing composite functions of the form

$$f(x) = g(x) + h(x)$$

where g is convex and differentiable, h is convex and nonsmooth but "simple"

For a lot of problems (i.e., functions h), we can recover the $O(1/\epsilon)$ rate of gradient descent with a simple algorithm, having important practical consequences

可分解函数

Suppose

$$f(x) = g(x) + h(x)$$

- g is convex, differentiable, $\mathrm{dom}(g) = \mathbb{R}^n$
- h is convex, not necessarily differentiable

If f were differentiable, then gradient descent update would be:

$$x^+ = x - t \cdot \nabla f(x)$$

Recall motivation: minimize quadratic approximation to f around x, replace $\nabla^2 f(x)$ by $\frac{1}{t} I$,

$$x^+ = \underset{z}{\mathrm{argmin}} \underbrace{f(x) + \nabla f(x)^T (z - x) + \frac{1}{2t} \|z - x\|_2^2}_{\bar{f}_t(z)}$$

典型例子：$\min_\beta \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$

In our case f is not differentiable, but $f = g + h$, g differentiable. Why don't we make quadratic approximation to g, leave h alone?

That is, update

$$
\begin{aligned}
x^+ &= \operatorname*{argmin}_z \ \bar{g}_t(z) + h(z) \\
&= \operatorname*{argmin}_z \ g(x) + \nabla g(x)^T (z - x) + \frac{1}{2t}\|z - x\|_2^2 + h(z) \\
&= \operatorname*{argmin}_z \ \frac{1}{2t}\big\|z - \big(x - t\nabla g(x)\big)\big\|_2^2 + h(z)
\end{aligned}
$$

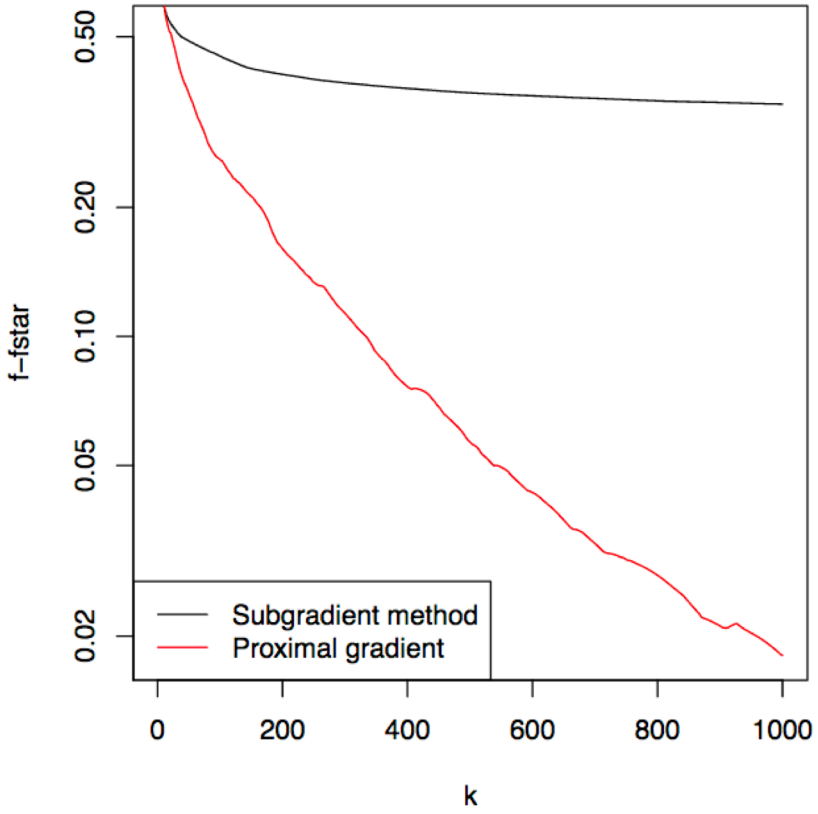
$\frac{1}{2t}\big\|z - \big(x - t\nabla g(x)\big)\big\|_2^2$ stay close to gradient update for g

$h(z)$ also make h small

上述二次逼近方式可导出一类更快的迭代算法

关于Lasso的对比

Example of proximal gradient (ISTA) vs. subgradient method convergence curves



Thank you for your attentions !