Proximal Gradient Descent Algorithms

Objectives

- derive proximal gradient algorithm for regularized least-squares problems
  - least-squares gradient descent
  - regularize
- apply to ridge regression

☆ The 5.1 section is to use GD way for solving LS problems

Here we are solving the (regularized-LS) in GD, just it.

Proximal gradient descent solves regularized least-squares problems

Problem: $\min_{\underline{w}} \|\underline{A}\underline{w} - \underline{d}\|_2^2 + \lambda r(\underline{w})$

$r(\underline{w})$: regularizer

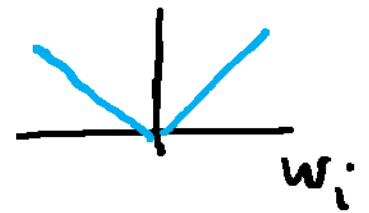
$\lambda > 0$: tuning parameter

Example Convex Regularizers

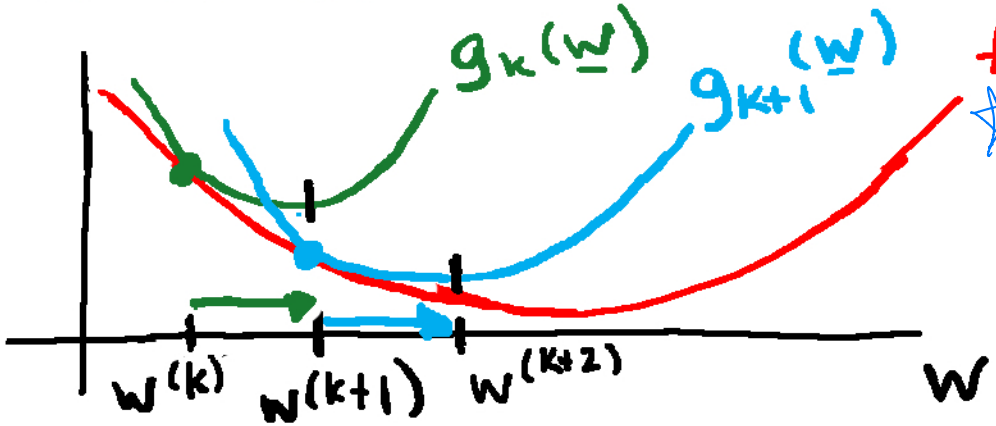
- Ridge (Tikhonov) $r(\underline{w}) = \|\underline{w}\|_2^2 = \sum_{i=1}^{n} w_i^2$
- LASSO (ℓ_1) $r(\underline{w}) = \|\underline{w}\|_1 = \sum_{i=1}^{n} |w_i|$

Not differentiable

not differ.



Proximal Gradient Descent Concept



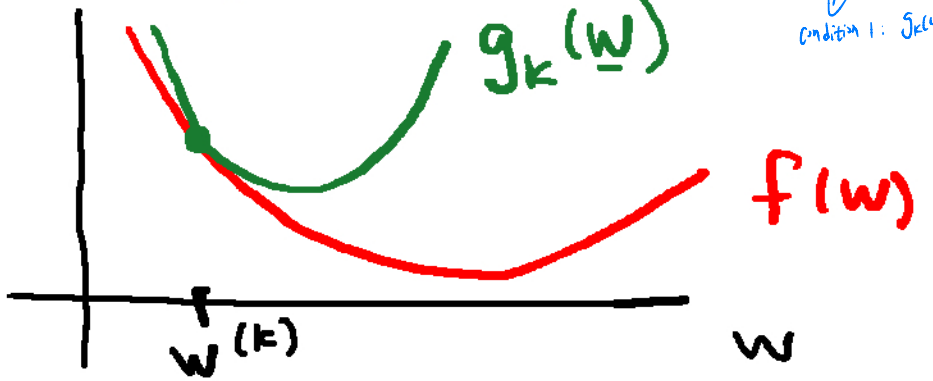
$f(\underline{w}) = \|\underline{A}\underline{w} - \underline{d}\|_2^2 + \lambda r(\underline{w})$

☆ 构建不近似的 $g_k(w)$ 作为 $f(w)$ 的 upper bound，descent 到 w_{k+1} → 构建新的 $g_{k+1}(w)$ → ....

- solve sequence of simpler problems
- simple for separable $r(\underline{w}) = \sum_i h_i(w_i)$ (sum of functions of w_i)

means there is no cross terms for $w_i w_j$, but just w_i

Find $g_k(\underline{w})$ so $f(\underline{w}) \le g_k(\underline{w})$, $g_k(\underline{w}^{(k)}) = f(\underline{w}^{(k)})$



condition 1: $g_k(\underline{w})$ is upper bound for $f(\underline{w})$

condition 2: At iteration k, the functions just touch

minimize $g_k(\underline{w}) \Rightarrow f(\underline{w})$ decreases

$f(\underline{w}^{(k+1)}) \le g_k(\underline{w}^{(k+1)}) \le g_k(\underline{w}^{(k)}) = f(\underline{w}^{(k)})$

$$f(\underline{w}) = \|\underline{d} - \underline{A}\underline{w}\|_2^2 + \lambda r(\underline{w})$$

$$= \|(\underline{d} - \underline{A}\underline{w}^{(k)}) + (\underline{A}\underline{w}^{(k)} - \underline{A}\underline{w})\|_2^2 + \lambda r(\underline{w})$$

Add a $\underline{A}\underline{w}^{(k)}$ here

$$f(\underline{w}) = \underbrace{\|\underline{d} - \underline{A}\underline{w}^{(k)}\|_2^2}_{c_k} + \underbrace{\|\underline{A}(\underline{w}^{(k)} - \underline{w})\|_2^2}_{\le \|\underline{A}\|_{op}^2 \|\underline{w}^{(k)} - \underline{w}\|_2^2} + 2\underbrace{(\underline{d} - \underline{A}\underline{w}^{(k)})^T \underline{A}}_{\underline{v}_k^T}(\underline{w}^{(k)} - \underline{w}) + \lambda r(\underline{w})$$

c_k: treat as constant, does not involve $\underline{w}$

$\le \|\underline{A}\|_{op}^2 \|\underline{w}^{(k)} - \underline{w}\|_2^2$: some standard property from theorem

$\underline{v}_k^T$: Does not involve $\underline{w}$ ($\underline{w}$ is the variable here)

Define step size $0 < \tau < 1/\|\underline{A}\|_{op}^2 \Rightarrow \frac{1}{\tau} > \|\underline{A}\|_{op}^2$

$$f(\underline{w}) \le g_k(\underline{w}) = c_k + \frac{1}{\tau}\|\underline{w}^{(k)} - \underline{w}\|_2^2 + 2\underline{v}_k^T(\underline{w}^{(k)} - \underline{w}) + \lambda r(\underline{w})$$

$g_k(\underline{w})$ is separable for $r(\underline{w})$ separable:

$$g_k(\underline{w}) = c_k + \sum_{i=1}^{m} q_i(w_i)$$

no $w_i w_j$ terms

可以拆解成 m 个互不相关的小问题 (1D)，而不是解决一个 m 维的大问题

Find $\underline{w}^{(k+1)} = \arg\min_{\underline{w}} g_k(\underline{w})$

$$g_k(\underline{w}) = c_k + \frac{1}{\tau}\|\underline{w}^{(k)} - \underline{w}\|_2^2 + 2\underline{v}_k^T(\underline{w}^{(k)} - \underline{w}) + \lambda r(\underline{w})$$

(multiply by τ)

$$\tau g_k(\underline{w}) = \tau c_k + (\underline{w}^{(k)} - \underline{w})^T(\underline{w}^{(k)} - \underline{w}) + 2\tau\underline{v}_k^T(\underline{w}^{(k)} - \underline{w}) + \lambda\tau r(\underline{w})$$

$$= \tau c_k - \tau^2\underline{v}_k^T\underline{v}_k + (\tau\underline{v}_k + (\underline{w}^{(k)} - \underline{w}))^T(\tau\underline{v}_k + (\underline{w}^{(k)} - \underline{w})) + \lambda\tau r(\underline{w})$$

$$\underline{z}^{(k)}$$

$$\underline{w}^{(k+1)} = \arg\min_{\underline{w}} \|\underline{z}^{(k)} - \underline{w}\|_2^2 + \lambda\tau r(\underline{w})$$

$$\underline{z}^{(k)} = \underline{w}^{(k)} + \tau\underline{v}_k$$

$$= \underline{w}^{(k)} + \tau\underline{A}^T(\underline{d} - \underline{A}\underline{w}^{(k)})$$

$$= \underline{w}^{(k)} - \tau\underline{A}^T(\underline{A}\underline{w}^{(k)} - \underline{d})$$

Least-squares gradient descent (Landweber)

Alternate LS gradient descent and regularization 5

交错执行依次 LS 梯度下降和正则化

initialize

$\underline{w}^{(0)} = \underline{0}$ (Initialize weights), $0 < \tau < \frac{1}{\|\underline{A}\|_{op}^2}$ (constrain the step size)

$\|A\|_{op} = \sigma_{max}(A) = \sqrt{\lambda_{max}(A^TA)}$ (singular value, eigenvalue)

LS gradient descent

$$\underline{z}^{(k)} = \underline{w}^{(k)} - \tau \underline{A}^T(\underline{A}\underline{w}^{(k)} - \underline{d})$$

regularize

$$\underline{w}^{(k+1)} = \arg\min_{\underline{w}} \|\underline{z}^{(k)} - \underline{w}\|_2^2 + \lambda\tau r(\underline{w})$$

check if converged

$$\text{if } \|\underline{w}^{(k+1)} - \underline{w}^{(k)}\| < \varepsilon \text{ stop}$$

Final recipe for prox algorithm

Regularization simple for $r(\underline{w})$ separable!

$$\text{if } r(\underline{w}) = \sum_{i=1}^{M} h_i(w_i)$$ 如果 r(w) 是可分离的

$$\underline{w}^{(k+1)} = \arg\min_{w_i, i=1,\dots M} \sum_{i=1}^{M} \left((z_i^{(k)} - w_i)^2 + \lambda\tau h_i(w_i) \right)$$

第三步可以替换成这个

M scalar minimizations

(M个一维最小化问题)

Example: Ridge Regression (Tikhonov)

$$f(\underline{w}) = \|\underline{d} - \underline{A}\underline{w}\|_2^2 + \lambda \|\underline{w}\|_2^2$$

LS gradient descent:

$$\underline{z}^{(k)} = \underline{w}^{(k)} - \tau \underline{A}^T (\underline{A}\underline{w}^{(k)} - \underline{d})$$

Regularization:

$$\underline{w}^{(k+1)} = \underset{w_i,\, i=1,\cdots M}{\operatorname{argmin}} \sum_{i=1}^{M} (z_i^{(k)} - w_i)^2 + \lambda\tau w_i^2$$

求导数让他=0，得到这时的 $w_i^{(k+1)}$

$$\Rightarrow w_i^{(k+1)} = \frac{1}{1+\lambda\tau} z_i^{(k)}$$

$$\underline{w}^{(k+1)} = \frac{1}{1+\lambda\tau} \underline{z}^{(k)}$$

"Shrink toward origin" 收缩