Proximal Gradient Descent Algorithms

Objectives

- derive proximal gradient algorithm for regularized least-squares problems
  - least-squares gradient descent
  - regularize

- apply to ridge regression

Proximal gradient descent solves regularized least-squares problems

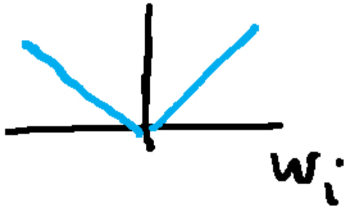
$$\min_{\underline{w}} \|\underline{A}\underline{w} - \underline{d}\|_2^2 + \lambda r(\underline{w})$$

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

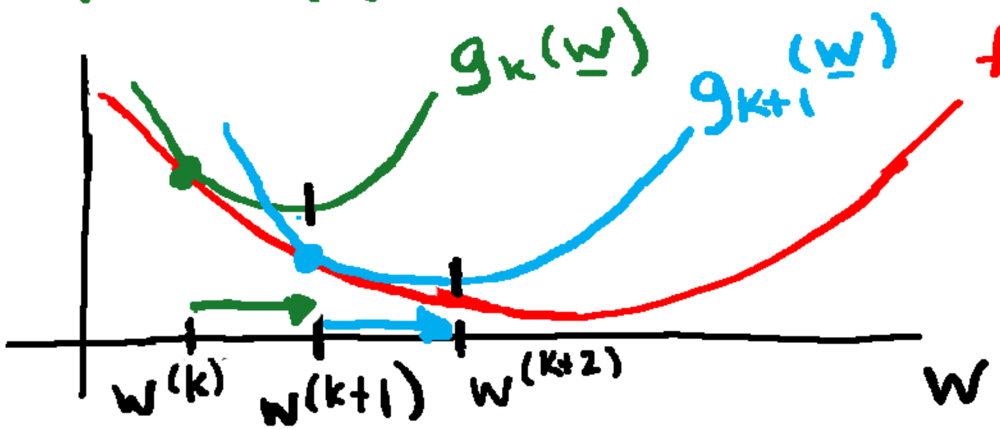
$\lambda > 0$: tuning parameter

Example Convex Regularizers

- Ridge (Tikhonov) $r(\underline{w}) = \|\underline{w}\|_2^2 = \sum_{i=1}^{M} w_i^2$
- LASSO (ℓ_1) $r(\underline{w}) = \|\underline{w}\|_1 = \sum_{i=1}^{M} |w_i|$ not differ.



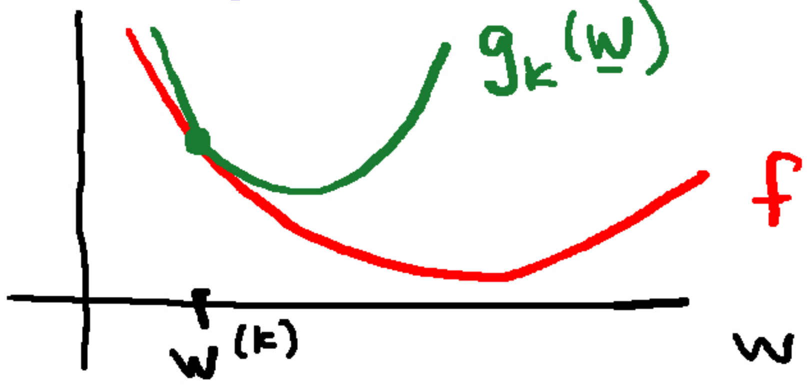
Proximal Gradient Descent Concept



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

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

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



minimize $g_k(w) \Rightarrow f(w)$ decreases

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

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

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

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: $\quad g_k(\underline{w}) = c_k + \sum_{i=1}^{M} g_i(w_i)$ $\quad$ no $w_i w_j$ terms

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

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

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

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

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

$$\begin{aligned}\underline{z}^{(k)} &= \underline{w}^{(k)} + \tau v_k \\ &= \underline{w}^{(k)} + \tau \underline{A}^T(\underline{d} - \underline{A}\underline{w}^{(k)}) \\ &= \underline{w}^{(k)} - \tau \underline{A}^T(\underline{A}\underline{w}^{(k)} - \underline{d})\end{aligned}$$

Least-squares gradient descent (Landweber)

Alternate LS gradient descent and regularization 5

$\underline{w}^{(0)} = \underline{0}$, $0 < \tau < \dfrac{1}{\|\underline{A}\|_{op}^2}$ initialize

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

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

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

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

if $r(\underline{w}) = \sum_{i=1}^{M} h_i(w_i)$

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

M scalar minimizations

Example: Ridge Regression (Tikhonov)

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

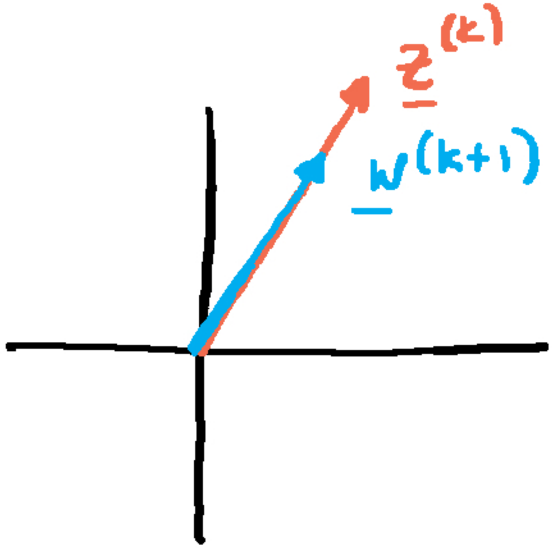
LS gradient descent:

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

Regularization:

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

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



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

"Shrink toward origin"