OPTIMIZATION FOR MACHINE LEARNING

November 18, 2024

Today: Regularized and sparse optimization (Part 1)

(1) Regularization

→ The basic optimization formulation of a machine learning task does not encode the full objective of that task

(Ex) $\{(a_i, y_i)\}_{i=1\dots m}$ (data) ⟶ $\underset{x \in \mathbb{R}^d}{\text{minimize}} \frac{1}{m}\sum_{i=1}^{m} f_i(x)$

"Empirical Risk Minimization"

Generalization ←
- What happens to the solution of the problem if the data is slightly perturbed?
- Can the solution get good results on new data?

Feature/Weight selection ←
- Is the solution interpretable? Is the resulting model simple?

⇒ How can those questions be encoded in the optimization formulation?

A: Regularization

Def: A regularized optimization problem has the form

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad f(x) + \lambda \Omega(x)$$

- $f(x)$: Data-fitting term, e.g. $f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$
- $\lambda > 0$: Regularization parameter (represents a trade-off between data fitting and regularization)
- $\Omega(x)$: Regularization term: represents properties that we would like the solution to have ← In general, does not depend on the data

- $\lambda \longrightarrow 0$: The problem eventually becomes equivalent to

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \; f(x)$$

- $\lambda \longrightarrow \infty$: The problem eventually becomes equivalent to

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \; \Omega(x)$$

(Ex) $\cdot \; \Omega(x) = \frac{1}{2}\|x\|^2 = \frac{1}{2}\|x\|_2^2 = \frac{1}{2}\sum_{i=1}^{d} x_i^2$

l_2 regularization / ridge regularization / Tychonov regularization

$\Rightarrow$ Used to reduce the sensitivity of the solution to perturbations of the data

$\Rightarrow \underset{x \in \mathbb{R}^d}{\text{minimize}} \; \frac{1}{2}\|x\|_2^2 \quad (\lambda \to \infty) \; : \Rightarrow$ solution is $x^* = 0_{\mathbb{R}^d}$

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \; f(x) + \frac{\lambda}{2}\|x\|_2^2 \;, \quad \lambda > 0$$

"convexifying effect"

[• Always strongly convex when f is convex, and in that case it has one unique minimum (whereas the original problem may have several)

$\Rightarrow$ Can be true even for nonconvex f provided λ is large enough

[Ex] $\underset{x\in\mathbb{R}^d}{\text{minimize}} \ \frac{1}{2m}\|Ax-y\|_2^2$

$d=2$, $m=2$, $A=\begin{bmatrix}1&0\\0&0\end{bmatrix}$ $y=\begin{bmatrix}1\\1\end{bmatrix}$

$\Rightarrow$ that problem is convex with infinitely many solutions

$\begin{bmatrix}1\\0\end{bmatrix}$ is typically the solution that we choose $\longrightarrow$ $\underset{x\in\mathbb{R}^2}{\text{argmin}} \ \frac{1}{4}\|Ax-y\|^2 = \left\{\begin{bmatrix}1\\t\end{bmatrix} \mid t\in\mathbb{R}\right\}$

Q] what happens if $A \longrightarrow A_\varepsilon = \begin{bmatrix}1+\varepsilon&0\\0&\varepsilon\end{bmatrix}$ for some $\varepsilon>0$

$$\underset{x\in\mathbb{R}^2}{\text{argmin}} \left\{\frac{1}{4}\|A_\varepsilon x-y\|^2\right\} = \left\{\begin{bmatrix}\frac{1}{1+\varepsilon}\\ \frac{1}{\varepsilon}\end{bmatrix}\right\}$$

$$\left\|\begin{bmatrix}\frac{1}{1+\varepsilon}\\ \frac{1}{\varepsilon}\end{bmatrix}\right\|^2 = \frac{1}{(1+\varepsilon)^2}+\frac{1}{\varepsilon^2} \longrightarrow \infty \quad \varepsilon\to 0$$

$\longrightarrow$ the solution (set) of $\underset{x\in\mathbb{R}^2}{\text{minimize}} \ \frac{1}{4}\|Ax-y\|^2$ is sensitive to perturbations of (the data) A.

$\longrightarrow$ Consider now

$A=\begin{bmatrix}1&0\\0&0\end{bmatrix}$ $y=\begin{bmatrix}1\\1\end{bmatrix}$ $\underset{x\in\mathbb{R}^2}{\text{minimize}} \ \frac{1}{4}\|Ax-y\|^2 + \frac{\lambda}{2}\|x\|^2$ $\lambda>0$

$\longrightarrow$ $\underset{x\in\mathbb{R}^2}{\text{argmin}} \left\{\underbrace{\frac{1}{4}\|Ax-y\|^2+\frac{\lambda}{2}\|x\|^2}_{f_\lambda(x)}\right\} = \left\{\begin{bmatrix}\frac{2}{2+\lambda}\\0\end{bmatrix}\right\}$

$$f_\lambda(x) = \frac{1}{4}\left\| \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\|^2 + \frac{\lambda}{2}(x_1^2 + x_2^2)$$

$$= \frac{1}{4}(x_1 - 1)^2 + \frac{1}{4} + \frac{\lambda}{2}x_1^2 + \frac{\lambda}{2}x_2^2$$

$$\nabla f_\lambda(x) = 0 \iff \begin{cases} \frac{1}{2}(x_1 - 1) + \lambda x_1 = 0 \\ \lambda x_2 = 0 \end{cases}$$

$$\iff \begin{cases} x_1 = \frac{1}{1+2\lambda} \\ x_2 = 0 \end{cases}$$

④ Unique solution (that is not a solution of the original problem $\underset{x \in \mathbb{R}^2}{\text{minimize}} \frac{1}{4}\|Ax - y\|^2$)

→ As $\lambda \to \infty$, the solution converges to $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$\underset{x \in \mathbb{R}^2}{\operatorname{argmin}} \frac{1}{2}\|x\|_2^2$

→ As $\lambda \to 0$, the solution converges to

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \in \underset{x \in \mathbb{R}^2}{\operatorname{argmin}} \left\{ \frac{1}{4}\|Ax - y\|^2 \right\}$$

↳ If we consider the perturbed problem

$$\underset{x \in \mathbb{R}^2}{\text{minimize}} \underbrace{\frac{1}{4}\|A_\varepsilon x - y\|^2 + \frac{\lambda}{2}\|x\|^2}_{f_{\lambda,\varepsilon}(x)}$$

$$\nabla f_{\lambda,\varepsilon}(x) = \frac{1}{2} A_\varepsilon^T (A_\varepsilon x - y) + \lambda x$$

$$= \frac{1}{2} \begin{bmatrix} 1+\varepsilon & 0 \\ 0 & \varepsilon \end{bmatrix} \begin{pmatrix} (1+\varepsilon) x_1 - 1 \\ \varepsilon x_2 - 1 \end{pmatrix} + \begin{bmatrix} \lambda x_1 \\ \lambda x_2 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{(1+\varepsilon)^2}{2} x_1 - \frac{1+\varepsilon}{2} + \lambda x_1 \\ \frac{\varepsilon^2 x_2}{2} - \frac{\varepsilon}{2} + \lambda x_2 \end{bmatrix}$$

$$\nabla f_{\lambda,\varepsilon}(x) = 0 \iff \begin{cases} x_1 = \frac{1+\varepsilon}{(1+\varepsilon)^2 + 2\lambda} \\ x_2 = \frac{\varepsilon}{\varepsilon^2 + 2\lambda} \end{cases}$$

$$\left(\lambda = 0 \quad \begin{bmatrix} \frac{1}{1+\varepsilon} \\ \frac{1}{\varepsilon} \end{bmatrix} \right)$$

$$\left\| \begin{bmatrix} \frac{1+\varepsilon}{(1+\varepsilon)^2 + 2\lambda} \\ \frac{\varepsilon}{\varepsilon^2 + 2\lambda} \end{bmatrix} \right\|^2 = \frac{(1+\varepsilon)^2}{\left((1+\varepsilon)^2 + 2\lambda\right)^2} + \frac{\varepsilon^2}{(\varepsilon^2 + 2\lambda)^2}$$

$$= O\left(\frac{1}{\lambda^2}\right)$$ for large values of λ and/or small values of ε

$$\left(\lambda \to \infty, \begin{bmatrix} \quad \end{bmatrix} \to \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right)$$

Ex) Sparsity-inducing regularization terms

Goal: Want a solution that is sparse, i.e. it has a significant number of zero coordinates.

⚠ Sparsity is a dimension-dependent notion

$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \in \mathbb{R}^2$

$\begin{bmatrix} * \\ \times \\ \times \\ 0 \\ ! \\ 0 \end{bmatrix}$ ↕ 1% nonzero coefficients

↕ 99% zero

$d = 10^9$

→ For linear models $a \longmapsto a^T x$ parameterized by x, finding a sparse linear model is a way to operate feature selection, i.e. to identify the most important features.

→ For overparameterized models (d parameters, n data points, $d >> n$),

there can exist infinitely models, some sparse and others dense (dense = all coefficients are nonzero)

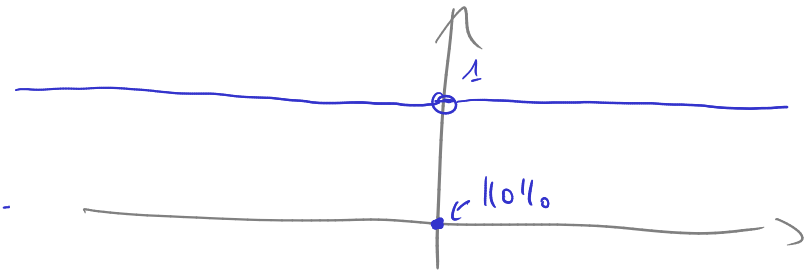
Regularizers that promote sparse solutions

- l_0 regularization

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \; f(x) + \lambda \|x\|_0 \quad \text{where} \quad \|x\|_0 = \sum_{j=1}^{d} \mathbb{1}(x_j \neq 0)$$

$$\mathbb{1}(t \neq 0) = \begin{cases} 1 & \text{if } t \neq 0 \\ 0 & \text{if } t = 0 \end{cases}$$

- $\|\cdot\|_0$ = "ℓ_0 norm" (but not a norm!)
- $\|x\|_0 \in \{0, 1, \dots, d\}$ for $x \in \mathbb{R}^d$
- $\|x\|_0$ is equal to the number of nonzero coordinates in x



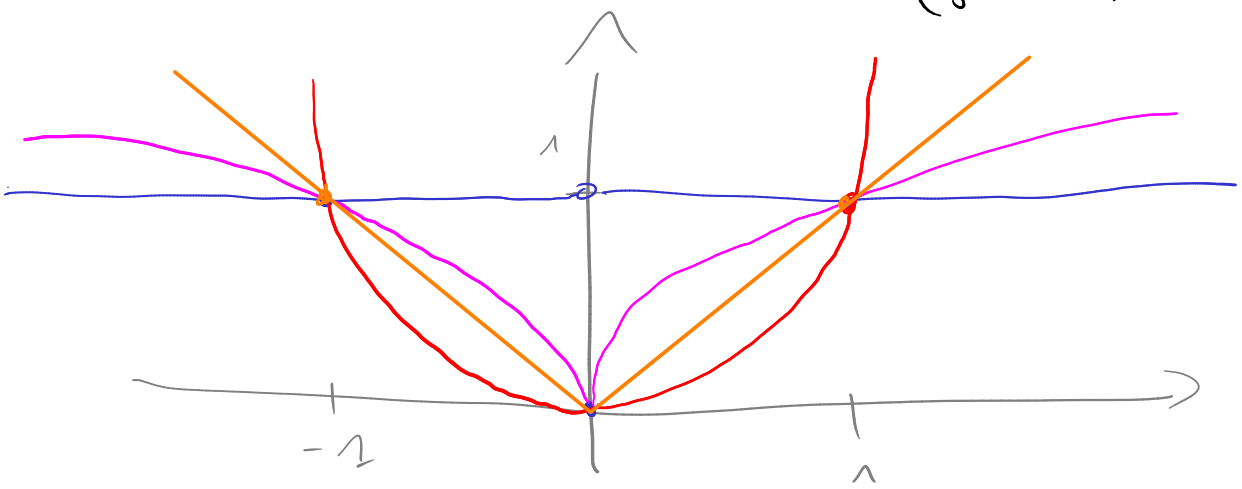
$d = 1$

$$\|x\|_0 = \begin{cases} 1 & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

⊖ $\|.\|_0$ is nonconvex, discontinuous so difficult to optimize!

→ Because the ℓ_0 norm is difficult to use in optimization, other regularization terms based on "ℓ_p norms" have been proposed

$$\forall\, p \in [0, +\infty], \quad \|x\|_p = \begin{cases} \sum_{j=1}^{d} \mathbb{1}(x_j \neq 0) & \text{if } p = 0 \\ \max_{1 \leq j \leq d} |x_j| & \text{if } p = \infty \\ \left(\sum_{j=1}^{d} |x_j|^p \right)^{1/p} & \text{if } 0 < p < \infty \end{cases}$$



— $p = 2$ $\|x\|_2^2$
— $p = 1$
— $p = 0$
— $p = 1/2$ $\|x\|_{1/2}^{1/2}$

→ ℓ_0 "norm" is the limit case of ℓ_p "norms" when $p \to 0$

⊖ For $p \in (0,1)$, $x \mapsto \|x\|_p^p$ is not convex nor a norm

⊖ For $p > 1$, $x \mapsto \|x\|_p^p$ is convex, easy to optimize because $\|x\|_p$ is a norm

But it is not the closest convex function to the ℓ_0 norm

→ $p = 1$ gives the closest convex function to the ℓ_0 norm, namely the ℓ_1 norm:

$$\|x\|_1 = \sum_{j=1}^{d} |x_j|$$

Convex, continuous, norm

Among all convex functions φ that satisfy $\varphi(x) \leq \|x\|_0 \ \forall x \in [-1,1]^d$ the ℓ_1 norm is the one that minimizes $\int_{[-1,1]^d} |\varphi(x) - \|x\|_0|$ over φ

⇒ ℓ_1 regularization is the dominant approach for enforcing sparsity.

(2) Proximal methods

Goal: Solve problems of the form

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \; f(x) + \lambda \Omega(x) \qquad f : \mathbb{R}^d \to \mathbb{R}, \quad \Omega : \mathbb{R}^d \to \mathbb{R}$$

↳ the methods that solve this problem while exploiting its structure are based on the concept of proximal operator

Def: The proximal operator of a function $h : \mathbb{R}^d \to \mathbb{R}$, denoted by $\text{prox}_h(\cdot)$, is a mapping defined by

$$\forall x \in \mathbb{R}^d, \quad \text{prox}_h(x) = \underset{u \in \mathbb{R}^d}{\text{argmin}} \left\{ h(u) + \frac{1}{2}\|u - x\|^2 \right\}$$

↳ Set-valued mapping in general: $\text{prox}_h(x)$ subset of $\mathbb{R}^d$

Example) $h(x) = \|x\|_0$

$$\text{prox}_h(x) = \left\{ v \in \mathbb{R}^d \mid \forall j = 1 \dots d, \right.$$

$$[v]_j = \begin{cases} 0 & \text{if } |x_j| < \sqrt{2} \\ x_j & \text{if } |x_j| > \sqrt{2} \\ \in \{0, x_j\} & \text{if } |x_j| = \sqrt{2} \end{cases} \left. \right\}$$

↳ If h is convex, the (proximal) prox operator of h can be

defined as a mapping from $\mathbb{R}^d \to \mathbb{R}^d$

- h convex $\Rightarrow h + \frac{1}{2}\|\cdot - x\|^2$ is strongly convex

$\Rightarrow \operatorname{argmin}_u \left\{ h(u) + \frac{1}{2}\|u-x\|^2 \right\}$ is a singleton

Examples:

- $h : x \mapsto 0$, $\operatorname{prox}_0(x) = x$ $\left(\operatorname{argmin}_u \frac{1}{2}\|u-x\|^2 = \{x\}\right)$

- $h : x \mapsto \frac{\lambda}{2}\|x\|_2^2$, $\operatorname{prox}_h(x) = \frac{x}{1+\lambda}$

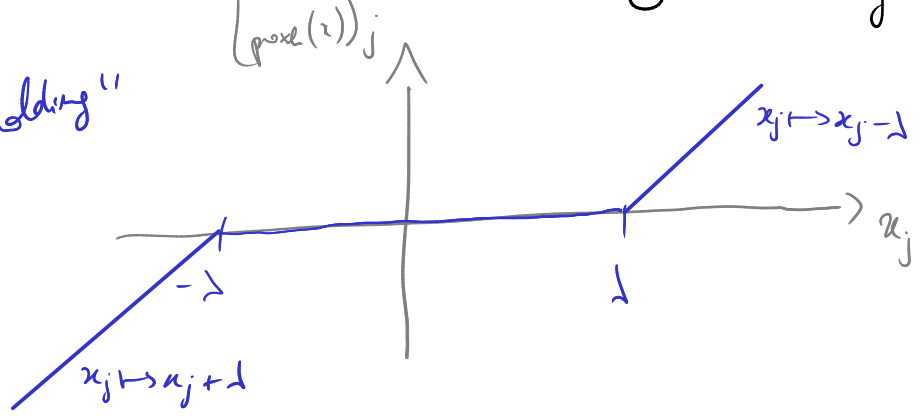
$\left(\lambda=1,\ \operatorname{argmin}_u \frac{1}{2}\|u-0\|^2 + \frac{1}{2}\|u-x\|^2 = \left\{\frac{x}{2}\right\}\right)$

- $h : x \mapsto \lambda\|x\|_1 = \lambda \sum_{j=1}^{d} |x_j|$

$\operatorname{prox}_h(x)$ is defined coordinate-wise by

$$\forall j = 1..d, \quad \left[\operatorname{prox}_h(x)\right]_j = \begin{cases} x_j - \lambda & \text{if } x_j > \lambda \\ x_j + \lambda & \text{if } x_j < -\lambda \\ 0 & \text{if } x_j \in [-\lambda, \lambda] \end{cases}$$

"soft-thresholding" operator



→ Prox operators are useful when they can be computed easily (often means that they are computed on a convex function and that the optimization problem defining the proximal operator can be solved explicitly)

→ Historically, the first method based on proximal operators was not assuming that the prox was easy to compute

This first method: Proximal point method

Goal: $\min_{x \in \mathbb{R}^d} h(x)$, $h : \mathbb{R}^d \to \mathbb{R}$ convex

Proximal point iteration ($x_0 \in \mathbb{R}^d$)

$$x_{k+1} = \text{prox}_{\alpha_k h}(x_k), \quad \alpha_k > 0$$

$$= \operatorname{argmin}_{x \in \mathbb{R}^d} \left\{ \alpha_k h(x) + \frac{1}{2} \|x - x_k\|_2^2 \right\}$$

$$= \operatorname{argmin}_{x \in \mathbb{R}^d} \left\{ h(x) + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right\}$$

Property: For any iteration k,

$$h(x_{k+1}) \leq h(x_k) - \frac{1}{2\alpha_k} \|x_{k+1} - x_k\|_2^2 < h(x_k)$$

if $x_{k+1} \neq x_k$

$$x_{k+1} = \operatorname{argmin}_x \left\{ h(x) + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right\}$$

$$h(x_{k+1}) + \frac{1}{2\alpha_k}\|x_{k+1} - x_k\|_2^2 \leq h(x_k) + \frac{1}{2\alpha_k}\|x_k - x_k\|_2^2 = h(x_k)$$

CV rate If $\alpha_k = \alpha > 0 \ \forall k \in \mathbb{N}$, then after K iterations,

$$h(x_K) - h(x_0) \leq \frac{\|x_0 - x^*\|^2}{2\alpha K}$$

where $x^* \in \operatorname{argmin}_{x \in \mathbb{R}^d} h(x)$

Drawback Each iteration to compute a proximal operator!

* When α_k is small, cost is similar to solving

$$\operatorname{minimize}_{x \in \mathbb{R}^d} \frac{1}{2}\|x - x_k\|^2 \quad \text{Easy!}$$

* When α_k is large, cost is similar to solving

$$\operatorname{minimize}_{x \in \mathbb{R}^d} h(x) \quad \text{Original problem!}$$

⟶ Iterations of the proximal point method can be as expensive as solving the original problem!

Special case: h C^1 and convex

In that case,

$$x_{k+1} = \text{prox}_{\alpha_k h}(x_k) \iff x_{k+1} = x_k - \alpha_k \nabla h(x_{k+1})$$

$\neq$ with gradient descent

Gradient evaluated at x_{k+1}

$$\text{prox}_{\alpha_k h}(x_k) = \underset{x \in \mathbb{R}^d}{\text{argmin}} \underbrace{\left\{ h(x) + \frac{1}{2\alpha_k}\|x - x_k\|_2^2 \right\}}_{\phi(x)}$$

ϕ is C^1 and strongly convex.

$$x_{k+1} = \underset{x \in \mathbb{R}^d}{\text{argmin}}\, \phi(x) \iff \nabla\phi(x_{k+1}) = 0$$

$$\iff \nabla h(x_{k+1}) + \frac{1}{\alpha_k}(x_{k+1} - x_k) = 0$$

$$\iff x_{k+1} = x_k - \alpha_k \nabla h(x_{k+1})$$

$\hookrightarrow$ Implicit definition of x_{k+1}, that can still be computed in special cases

Example: $h(x) = \frac{1}{2m}\|Ax - y\|_2^2$ $\quad A \in \mathbb{R}^{m \times d}$, $b \in \mathbb{R}^m$

$$\nabla h(x) = \frac{1}{m} A^T(Ax - y)$$

$$x_{k+1} = x_k - \alpha_k \nabla h(x_{k+1}) = x_k - \frac{\alpha_k}{m} A^T A x_{k+1} + \frac{\alpha_k}{m} A^T y$$

$$\iff \left[I + \frac{\alpha_k}{m} A^T A \right] x_{k+1} = x_k + \frac{\alpha_k}{m} A^T y$$

$$\Leftrightarrow \quad x_{k+1} = \left[I + \frac{\alpha_k}{n} A^T A \right]^{-1} \left(x_k + \frac{\alpha_k}{n} A^T y \right)$$

Proximal gradient method

↳ Dedicated to problems of the form

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad f(x) + \lambda \Omega(x)$$

$f : \mathbb{R}^d \to \mathbb{R}$ C^1

$\Omega : \mathbb{R}^d \to \mathbb{R}$ convex

$\lambda > 0$

Idea:

- Optimizing f alone could be done using gradient descent
- Optimizing Ω alone ——— the proximal point method

Proximal gradient iteration

$$x_{k+1} = \text{prox}_{\alpha_k \lambda \Omega} \left(x_k - \alpha_k \nabla f(x_k) \right) \qquad \alpha_k > 0$$

$$= \underset{x \in \mathbb{R}^d}{\text{argmin}} \left\{ f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2\alpha_k} \|x - x_k\|^2 + \lambda \Omega(x) \right\}$$

$$\text{prox}_{\alpha_k \lambda \Omega} \left(x_k - \alpha_k \nabla f(x_k) \right) = \underset{x \in \mathbb{R}^d}{\text{argmin}} \left\{ \alpha_k \lambda \Omega(x) + \frac{1}{2} \|x - (x_k - \alpha_k \nabla f(x_k))\|^2 \right\}$$

$\times \frac{1}{\alpha_k}$ ↳

$$= \underset{x \in \mathbb{R}^d}{\text{argmin}} \left\{ \frac{1}{2\alpha_k} \|x - x_k + \alpha_k \nabla f(x_k)\|^2 + \lambda \Omega(x) \right\}$$

$$= \operatorname{argmin}_{x \in \mathbb{R}^d} \left\{ \frac{1}{2\alpha_k} \|x - x_k\|^2 + \frac{1}{\alpha_k}(x - x_k)^T(\alpha_k \nabla f(x_k)) + \frac{1}{\alpha_k}\|\alpha_k \nabla f(x_k)\|^2 + \lambda \Omega(x) \right\}$$

$$= \operatorname{argmin}_{x \in \mathbb{R}^d} \left\{ \frac{1}{\alpha_k}\|\alpha_k \nabla f(x_k)\|^2 + \nabla f(x_k)^T(x - x_k) + \frac{1}{2\alpha_k}\|x - x_k\|^2 + \lambda \Omega(x) \right\}$$

↓ constant w.r.t. x

$$= \operatorname{argmin}_{x \in \mathbb{R}^d} \left\{ f(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{1}{2\alpha_k}\|x - x_k\|^2 + \lambda \Omega(x) \right\}$$

$f(x_k)$: another constant w.r.t. x

$f(x_k) + \nabla f(x_k)^T(x - x_k)$: approximation of $f(x)$ around x_k

$\frac{1}{2\alpha_k}\|x - x_k\|^2$: proximal term, penalizes x's that are far from x_k

$\lambda \Omega(x)$: regularization term (untouched)

$f \equiv 0$: $x_{k+1} = \operatorname{prox}_{\alpha_k \lambda \Omega}(x_k - \alpha_k \nabla f(x_k)) = \operatorname{prox}_{\alpha_k \lambda \Omega}(x_k)$

Proximal point applied to $\lambda\Omega$

$\Omega \equiv 0$: $x_{k+1} = \operatorname{prox}_{\alpha_k \lambda \Omega}(x_k - \alpha_k \nabla f(x_k)) = x_k - \alpha_k \nabla f(x_k)$

GD applied to f

(+) Uses the problem structure (convexity of Ω, C^1 nature of f)

(+) Applies when f is nonconvex

(−) Requires to compute a prox for the function Ω

Two key examples

- ℓ_2 regularization

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad f(x) + \frac{\lambda}{2}\|x\|_2^2 \qquad \lambda > 0$$

$f \in C^1$

PG

$$x_{k+1} = \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \underbrace{f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2\alpha_k}\|x - x_k\|^2 + \frac{\lambda}{2}\|x\|^2}_{\varphi(x)} \right\}$$

$$\nabla \varphi(x) = 0 \iff \nabla f(x_k) + \frac{1}{\alpha_k}(x - x_k) + \lambda x = 0$$

$$\iff \frac{1+\lambda\alpha_k}{\alpha_k} x = \frac{1}{\alpha_k} x_k - \nabla f(x_k)$$

$$\iff x = \frac{1}{1+\lambda\alpha_k} x_k - \frac{\alpha_k}{1+\lambda\alpha_k} \nabla f(x_k)$$

$\lambda = 0 \qquad x = x_k - \alpha_k \nabla f(x_k)$

PG for ℓ_2 regularization

$$x_{k+1} = \underbrace{\frac{1}{1+\lambda\alpha_k}}_{< 1} x_k - \underbrace{\frac{\alpha_k}{1+\lambda\alpha_k}}_{< 1 \text{ for } \lambda \text{ sufficiently large}} \nabla f(x_k)$$

$\Rightarrow$ this method performs weight decay and gradient decay for λ sufficiently large

GD with weight decay: GD applied to $\min_{x \in \mathbb{R}^d} f(x) + \frac{\lambda}{2}\|x\|_2^2$

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) - \lambda \alpha_k x_k$$

$$= \underbrace{(1 - \lambda \alpha_k)}_{< 1 \text{ for } \alpha_k < \frac{1}{\lambda}} x_k - \alpha_k \nabla f(x_k)$$

NB: Proximal point method applied to $\min_{x \in \mathbb{R}^d} f(x) + \frac{\lambda}{2}\|x\|_2^2$

$$x_{k+1} = x_k - \alpha_k \nabla \left(f + \frac{\lambda}{2}\|\cdot\|_2^2\right)(x_{k+1})$$

$$= x_k - \alpha_k \nabla f(x_{k+1}) - \lambda \alpha_k x_{k+1}$$

$$[1 + \lambda \alpha_k]\, x_{k+1} = x_k - \alpha_k \nabla f(x_{k+1})$$

$$x_{k+1} = \frac{1}{1+\lambda \alpha_k} x_k - \frac{\alpha_k}{1+\lambda \alpha_k} \nabla f(x_{k+1})$$

Takeaways

- Regularization $\equiv$ Pushing for certain properties of the solution
- ℓ_2, (ℓ_0), ℓ_1 → Tuesday
- Proximal operator, proximal point, proximal gradient

Exercise:

(P) minimize $\|x\|_1 + \frac{1}{2\alpha}\|x-u\|_2^2$ for some $u \in \mathbb{R}^d$, $\alpha > 0$
$x \in \mathbb{R}^d$

1) Show that the solution of (P) is given by the proximal operator of a certain function

2) Write down the iteration of proximal gradient applied to (P) with $f(x) = \frac{1}{2}\|x-u\|_2^2$, $\Omega(x) = \|x\|_1$ and $\lambda = \alpha$.

(Explicit form if you can)