OPTIMIZATION FOR MACHINE LEARNING

Regularized, large-scale and distributed optimization

November 9, 2023

REGULARIZATION AND PROXIMAL METHODS

① Regularization

↳ Most data science tasks are formulated in an incomplete way as optimization problem

Ex) * Want the model that is learned through optimization to generalize to unseen data

* Would like models that are interpretable, ideally simple

⟹ In general, these properties are hard to encode in an optimization formulation

↳ Typical learning problem:

regularization parameter $\lambda > 0$ gives more or less weight to regularization

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

"data-fitting term" $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$

"regularization term" $\Omega : \mathbb{R}^d \to \mathbb{R}$ typically does not depend on data

$\lambda = 0$: $\underset{x \in \mathbb{R}^d}{\text{minimize}} \ f(x)$

$\lambda \to +\infty$ The problem essentially becomes

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

: That problem does not depend on data at all

Ω represents properties that we would like the solution to satisfy and λ represents the tradeoff between data fitting and regularization

Ex) • ℓ_2 regression / ridge regression: $\Omega(x) = \frac{1}{2}\|x\|^2$
(aka Tychonov regularization)

$$= \frac{1}{2} x^T x$$

$$= \frac{1}{2} \sum_{j=1}^{d} x_j^2$$

→ Leads to solutions that are less sensitive to variations in the data

$\left(\lambda \to \infty : \quad \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{2}\|x\|^2 = \{0\} \right)$

• ℓ_1 regression / LASSO : $\Omega(x) = \|x\|_1 = \sum_{j=1}^{d} |x_j|$

→ Leads to solutions that are sparse (a significant amount of zero coefficients)

• Variations on ℓ_1 and ℓ_2:

↳ Elastic net $\quad \Omega(x) = \|x\|_1 + \frac{\mu}{2}\|x\|^2$

$\mu > 0$

↳ Group LASSO: $\quad \Omega(x) = \sum_{g \in \mathcal{G}} \|x_g\|$

$$x = \begin{bmatrix} x_{g_1} \\ \vdots \\ x_{g_m} \end{bmatrix} \quad \mathcal{G} = \{g_1, \dots, g_m\}$$

• Constraint $x \in X \subseteq \mathbb{R}^d$

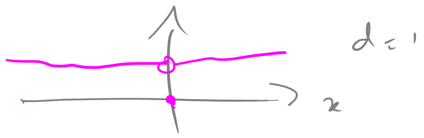
$$\Omega(x) = \begin{cases} 0 & \text{if } x \in X \\ \infty & \text{otherwise} \end{cases}$$

$$\left(f(x) + \overset{\lambda > 0}{\lambda} \Omega(x) = \begin{cases} f(x) & \text{if } x \in X \\ \infty & \text{otherwise} \end{cases} \right)$$

$\rightarrow \Omega$ can be of many different forms (convex/non-convex, smooth C^1 / nonsmooth, continuous/discontinuous)
$\|\cdot\|^2$ $\qquad$ $\|\cdot\|_1$

Ex) $\Omega(x) = \|x\|_0$ ("ℓ_0-norm"),

$\|x\|_0 = |\{j \in \{1,\dots,d\},\ x_j \neq 0\}|$



What class of algorithms can we use to solve the resulting optimization problems?
$\Rightarrow$ Proximal methods

(2) Proximal operators

Def: Let $h : \mathbb{R}^d \to \mathbb{R}$ that is closed ($h(\mathbb{R}^d)$ is a closed set), proper (h takes at least one finite value) and convex.

The proximal operator of h, denoted by $\text{prox}_h(\cdot)$, is the function from $\mathbb{R}^d$ to $\mathbb{R}^d$ 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 in $\mathbb{R}^d$

strongly convex function (unique minimum)

argmin$\{\}$ is a singleton

$\text{prox}_h(x)$ is well-defined as the unique solution to a strongly convex optimization problem

Ex) $\text{prox}_0(x) = \underset{u \in \mathbb{R}^d}{\text{argmin}} \left\{ 0 + \frac{1}{2}\|u - x\|^2 \right\} = x$

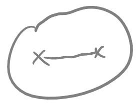
$\text{prox}_{\frac{1}{2}\|\cdot\|^2}(x) = \underset{u \in \mathbb{R}^d}{\text{argmin}} \left\{ \frac{1}{2}\|u\|^2 + \frac{1}{2}\|u - x\|^2 \right\} = \frac{x}{2}$

$\forall \lambda > 0 \qquad \mathrm{prox}_{\frac{\lambda}{2}\|\cdot\|^2}(x) = \dfrac{x}{1+\lambda} \xrightarrow[\lambda \to \infty]{} 0$

. $\mathrm{prox}_{\lambda\|\cdot\|_1}(x)$ is defined coordinatewise by

$$\forall j = 1..d, \quad [\mathrm{prox}_{\lambda\|\cdot\|_1}(x)]_j = \begin{cases} x_j - \lambda & \text{if } x_j > \lambda \\ x_j + \lambda & \text{if } x_j < -\lambda \\ 0 & \text{if } x_j \in [-\lambda, \lambda] \end{cases}$$

. $h : x \longmapsto \begin{cases} 0 & \text{if } x \in \mathcal{X} \\ \infty & \text{otherwise} \end{cases}$ with $\mathcal{X}$ convex set in $\mathbb{R}^d$



$$\mathrm{prox}_h(x) = \begin{cases} x & \text{if } x \in \mathcal{X} \\ \text{projection of } x \text{ onto } \mathcal{X} & \text{if } x \notin \mathcal{X} \end{cases}$$

$$= \underset{u \in \mathbb{R}^d}{\mathrm{argmin}} \left\{ \frac{1}{2}\|u - x\|^2 \text{ s.t. } x \in \mathcal{X} \right\}$$

↳ Proximal operators are interesting when

i) they are uniquely defined (true when h is convex, sometimes true for nonconvex or even discontinuous h !)

ii) they can be computed easily

⟹ With these two properties, you can consider using these operators in optimization algorithms

(3) Proximal point method

Problem: $\underset{x \in \mathbb{R}^d}{\text{minimize}} \; h(x)$ $\qquad$ h convex function $h : \mathbb{R}^d \to \mathbb{R}$

Proximal point iteration ($k \in \mathbb{N}$)

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

where $\alpha_k > 0$

Property: $\forall k \in \mathbb{N}$

x_{k+1} minimum of $h + \frac{1}{2\alpha_k}\|\cdot - x_k\|^2$

$$h(x_{k+1}) + \frac{1}{2\alpha_k}\|x_{k+1} - x_k\|^2 \leq h(x_k) + \overbrace{\frac{1}{2\alpha_k}\|x_k - x_k\|^2}^{=0}$$

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

guaranteed decrease when $x_{k+1} \neq x_k$

Consequences:
→ Can prove convergence rates for that method

$\forall K \geq 1, \quad h(x_K) - \min_{x \in \mathbb{R}^d} h(x) \leq O\left(\frac{1}{K}\right)$

→ Per-iteration cost: Compute a proximal operator, i.e. solve an optimization problem

(+) The subproblems $x_{k+1} = \mathrm{prox}_{\alpha_k h}(x_k)$ are strongly convex even if h is convex not strongly convex

(−) Depending on h, computing the "prox" (proximal operator) can be as expensive as solving the original problem

Special case $h \in C^1$ (and convex)

$$\underset{x\in\mathbb{R}^d}{\operatorname{argmin}} \Big\{ \underbrace{\underbrace{h(x)}_{C^1} + \underbrace{\frac{1}{2\alpha_k}\|x-x_k\|^2}_{C^1}}_{\text{strongly convex}} \Big\}$$

is the singleton containing the unique solution to

$$(*) \quad \nabla\Big(h + \frac{1}{2\alpha_k}\|\cdot - x_k\|^2\Big)(x) = 0$$

$$\nabla\Big(\frac{1}{2}\|\cdot - a\|^2\Big)(x) = x - a$$

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

$$\iff x = x_k - \alpha_k \nabla h(x)$$

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

$\to$ Implicit method (no close form for x_{k+1})

Compare the iteration with GD: $x_{k+1} = \underbrace{x_k - \alpha_k \nabla h(x_k)}$

Explicit calculation of x_{k+1}

Special sub-case

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

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

Gradient descent:

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

Proximal point:

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

$$\underbrace{\Big[I + \underbrace{\frac{\alpha_k}{m}A^TA}_{\succeq 0}\Big]}_{\text{invertible}} x_{k+1} = x_k + \frac{\alpha_k}{m}A^Ty$$

$$x_{k+1} = \underbrace{\Big[I + \frac{\alpha_k}{m}A^TA\Big]^{-1}\Big(x_k + \frac{\alpha_k}{m}A^Ty\Big)}$$

explicit formula for x_{k+1}

$\Rightarrow$ Here each iteration of the proximal point method requires to solve a linear system

4) Proximal gradient method

↳ Consider again $\underset{x \in \mathbb{R}^d}{\text{minimize}} \ \overbrace{f(x) + \lambda \Omega(x)}^{h(x)}$

where f is C^1 (possibly nonconvex)

and Ω is convex

⇒ Instead of applying the proximal point method to h, we would like to exploit the structure of h and in particular that of f

Proximal gradient iteration

↳ Starting from x_k, compute $\nabla f(x_k)$. and $\alpha_k > 0$

↳ Compute

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

regularization expressed at x

good approximation of f near x_k

proximal term that penalizes x away from x_k

Theorem: The proximal gradient iteration corresponds to

$$x_{k+1} = \operatorname{prox}_{\alpha_k \lambda \Omega}\Big(\underbrace{x_k - \alpha_k \nabla f(x_k)}_{\text{Gradient step with stepsize } \alpha_k} \Big)$$

NB: if $\Omega \equiv 0$ (no regularization) $x_{k+1} = \operatorname{prox}_0 (x_k - \alpha_k \nabla f(x_k)) = x_k - \alpha_k \nabla f(x_k)$

Proximal gradient without regularization = Gradient descent

Proof: $\text{prox}_{\alpha_k \lambda \Omega}\left(x_k - \alpha_k \nabla f(x_k)\right)$

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

$\alpha_k > 0$

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

$\|a+b\|^2 = \|a\|^2 + 2a^Tb + \|b\|^2$

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

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

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

$+f(x_k)$

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

↳ Proximal gradient combines gradient descent on f with a prox on (a multiple of) Ω

- Works with any convex Ω (because such a function is "proximable", i.e. its proximal operator is well-defined) and for some nonconvex Ω
- Useful when the prox operator for Ω is easy to compute

Focus: $\Omega(x) = \frac{1}{2}\|x\|^2$

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

$\lambda \to \infty$: $\min \frac{1}{2}\|x\|^2 \to x^* = 0$

$\lambda > 0$: ?

$\lambda \to 0$: $\underset{x \in \mathbb{R}^d}{\text{minimize}}\ f(x)$

Proximal gradient iteration

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

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

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

shrinking the coefficients of x_k

stepsize $< \alpha_k$

$\Rightarrow$ Similar to weight decay in stochastic gradient / gradient descent

$\hookrightarrow$ For that problem, since $x \mapsto \frac{1}{2}\|x\|^2$, we could also apply GD!

$$\nabla\left(f + \frac{\lambda}{2}\|\cdot\|^2\right)(x) = \nabla f(x) + \lambda x$$

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

$$= (1 - \lambda\alpha_k)x_k - \alpha_k \nabla f(x_k)$$

"Weight decay"

$\hookrightarrow$ because of this term, $\|x_{k+1}\|$ might be large even if $\lambda \gg 1$

l_2 regularization + GD $\Rightarrow$ weight decay

l_2 ——————— + PG $\Rightarrow$ ——————— + gradient decay

As $\lambda \to \infty$, the iteration gets closer to $x_{k+1} = 0$

But for $\lambda > 0$, the components of the iterates will decrease in a smooth fashion (they will all converge to 0 in the same way)

$\hookrightarrow$ If $x(\lambda)$ is a solution of $\underset{x \in \mathbb{R}^d}{\text{minimize}} \; f(x) + \frac{\lambda}{2} \|x\|^2$, we can show that

$$\lambda_2 \geq \lambda_1 \quad \Rightarrow \|x(\lambda_2)\|^2 \leq \|x(\lambda_1)\|^2$$

$\Rightarrow$ Regularization reduces the norm of the solution, prevents from very large values

$\hookrightarrow$ l_2 regularization reduces the variance with respect to the data

Suppose that we observe $b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $A = \begin{bmatrix} 1+\varepsilon & 0 \\ 0 & \varepsilon \end{bmatrix}$

ε : noise > 0

Linear regression on (A, b)

$$\underset{x \in \mathbb{R}^2}{\text{minimize}} \; \frac{1}{4} \|Ax - b\|^2$$

solution given by $A^T A x - A^T b = 0$

$$A^T A = \begin{bmatrix} (1+\varepsilon)^2 & 0 \\ 0 & \varepsilon^2 \end{bmatrix} \qquad A^T b = \begin{bmatrix} 1+\varepsilon \\ \varepsilon \end{bmatrix}$$

$$\underbrace{x(0)}_{\text{solution without regularization}} = \begin{bmatrix} \frac{1}{1+\varepsilon} \\ \frac{1}{\varepsilon} \end{bmatrix}$$

For small noise, $\|x(0)\| = O\left(\frac{1}{\varepsilon}\right)$ blows up!

$Ax(0) = b$ but in terms of the real data

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{1+\varepsilon} \\ 1/\varepsilon \end{bmatrix} = \begin{bmatrix} \frac{1}{1+\varepsilon} \\ 0 \end{bmatrix} \neq \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

As $\varepsilon \to \infty$, the value of the solution gets worse in terms of fitting the noiseless data

$\hookrightarrow$ If we consider the problem without noise

$$\underset{x \in \mathbb{R}^2}{\text{minimize}} \ \frac{1}{2} \left\| \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} x - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\|^2,$$

This is a convex problem with infinitely many solutions

$\Rightarrow$ which one should we choose?

With ℓ_2 regularization

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \ \frac{1}{2} \|Ax - b\|^2 + \frac{\lambda}{2} \|x\|^2$$

$\Rightarrow$ strongly convex problem

$\Rightarrow$ Solution $\quad x(\lambda) = \begin{bmatrix} (1+\varepsilon)^2+\lambda & 0 \\ 0 & \varepsilon^2+\lambda \end{bmatrix}^{-1} \begin{bmatrix} 1+\varepsilon \\ \varepsilon \end{bmatrix}$

$$= \begin{bmatrix} \dfrac{1+\varepsilon}{(1+\varepsilon)^2+\lambda} \\ \dfrac{\varepsilon}{\varepsilon^2+\lambda} \end{bmatrix}$$

$$\|x(\lambda)\| \underset{\lambda\to\infty}{\longrightarrow} 0$$

$\varepsilon = 0$ (actually no noise) $= \begin{bmatrix} \frac{1}{1+\lambda} \\ 0 \end{bmatrix} \underset{\lambda\to 0}{\longrightarrow} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$\varepsilon \to \infty \quad x(\lambda) \to \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ in a smooth way

"Minimum norm solution of the problem without noise"

Several reasons to use ℓ_2 regularization

- Want to minimize f convex but there are multiple solutions
  $\Rightarrow$ with ℓ_2 regularizat°, get a unique solution!
  $\Rightarrow$ what proximal methods do!

- Want to reduce the dependency of the solution on the data defining f
  $\Rightarrow$ A way of tackling overfitting

-

$\hookrightarrow$ Key: Tradeoff between data fitting and regularization
$\Rightarrow$ choice of λ :