Optimization for Data Science
ETH Zürich, FS 2023 261-5110-00L

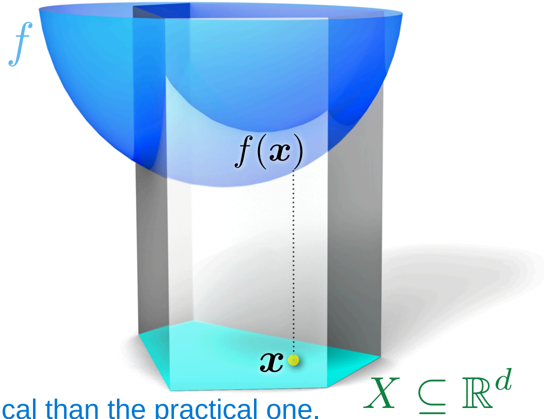
Lecture 4: Projected Gradient Descent

Bernd Gärtner
Niao He

https://www.ti.inf.ethz.ch/ew/courses/ODS23/index.html

March 13, 2023

Constrained Optimization



Constrained Optimization Problem

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in X \end{array}$$

this is more theoritical than the practical one.

- Lecture 3: $X = \mathbb{R}^d$ (unconstrained optimization)
- This lecture: $X \subsetneq \mathbb{R}^d$ (closed convex set)

Example: Master's Admission

- CS department of a well known Swiss university is admitting top international students to its MSc program, in a competitive application process.
- Applicants are submitting various documents (GPA, TOEFL test score, GRE test scores, reference letters,...)
- Admission committee would like to compute a (rough) forecast of the applicant's performance in the MSc program, based on the submitted documents.
- Data on the actual performance of students admitted in the past is available.
- In the following (**made-up toy**) example: consider GPA and TOEFL only...
- ...as predictors for GGPA (graduation grade point average; final grade obtained in MSc program)
- Real GGPA prediction (using machine learning techniques) has been investigated in a doctoral thesis at ETH [Zim16].

Example: Master's Admission

- $0.0 \leq$ GPA ≤ 4.0 (admission starts from 3.5)
- $0 \leq$ TOEFL ≤ 120 (admission starts from 100)
- $1.0 \leq$ GGPA ≤ 6.0 (Swiss grading scale)
- Historical data from students admitted in the past:

x_1 (GPA)	x_2 (TOEFL)	y (GGPA)
3.52	100	3.92
3.66	109	4.34
3.76	113	4.80
3.74	100	4.67
3.93	100	5.52
3.88	115	5.44
3.77	115	5.04
3.66	107	4.73
3.87	106	5.03
3.84	107	5.06

Master's Admission: hypothesis class and loss function

Assumption: linear model!

$$\mathrm{GGPA} \approx w_0 + w_1 \cdot \mathrm{GPA} + w_2 \cdot \mathrm{TOEFL}$$

for unknown weights w_0, w_1, w_2.

- Hypothesis class $\mathcal{H} = \mathbb{R}^3 = \{(w_0, w_1, w_2)\}$

Approach: minimize least squares error over the historical data.

- Loss function $\ell(\mathbf{w}, (\mathbf{x}, y)) = (w_0 + w_1 x_1 + w_2 x_2 - y)^2$

Empirical risk minimizer weights $w_1^\star, w_2^\star$ should tell us how indicative GPA and TOEFL are for the GGPA (large weight $\approx$ high influence).

- Relevant GPA scores span a range of 0.5.
- Relevant TOEFL scores span a range of 20.
- Normalize first so that w_1, w_2 can be compared.
- Details in Section 2.6.2.

Master's Admission: Normalized data

x_1 (GPA)	x_2 (TOEFL)	y (GGPA)
-2.04	-1.28	-0.94
-0.88	0.32	-0.52
-0.05	1.03	-0.05
-0.16	-1.28	-0.18
1.42	-1.28	0.67
1.02	1.39	0.59
0.06	1.39	0.19
-0.88	-0.04	-0.12
0.89	-0.21	0.17
0.62	-0.04	0.21

Empirical risk $\ell_{10}(w_1, w_2)$ ($w_0 = 0$ after normalization):

$$f(w_1, w_2) = \sum_{i=1}^{10}(w_1 x_{i1} + w_2 x_{i2} - y_i)^2 \approx 10w_1^2 + 10w_2^2 + 1.99 w_1 w_2 - 8.7 w_1 - 2.79 w_2 + 2.09.$$

Master's Admission: Empirical risk minimization

Optimal solution: $(w_1^\star, w_2^\star) \approx (0.43, 0.097)$

Under our hypothesis (linear model), we therefore expect $y_i \approx y_i^\star = 0.43x_{i1} + 0.097x_{i2}$

x_{i1}	x_{i2}	y_i	$y_i^\star$	$z_i^\star$
-2.04	-1.28	-0.94	-1.00	-0.87
-0.88	0.32	-0.52	-0.35	-0.37
-0.05	1.03	-0.05	0.08	-0.02
-0.16	-1.28	-0.18	-0.19	-0.07
1.42	-1.28	0.67	0.49	0.61
1.02	1.39	0.59	0.57	0.44
0.06	1.39	0.19	0.16	0.03
-0.88	-0.04	-0.12	-0.38	-0.37
0.89	-0.21	0.17	0.36	0.38
0.62	-0.04	0.21	0.26	0.27

Not too bad: Low empirical risk on the training data... even if we only use the GPA to predict the GGPA:

$$y_i \approx z_i^\star = 0.43x_{i1}$$

Master's Admission: Expected risk minimization?

Known problems with least squares:

there are factors that will not influence the final result, which should be dropped.

- Likely to overfit.
- "Unimportant" variables should have weight 0, but they typically don't

Subset selection heuristics: drop variables with seemingly "small" contribution (various methods to decide what "small" means, and how many to drop)

Best subset selection: solve least squares subject to an additional constraint that there are at most k nonzero weights (NP-hard; various k might have to be tried)

Regularization: solve least squares subject to an additional constraint that $\mathbf{w}$ has small norm. (As norms are convex, we get a convex feasible set X)

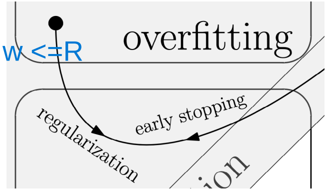
get a better result via minimizing a constructed constraint.

LASSO: popular regularization method with some favorable statistical properties: considers the 1-norm of $\mathbf{w}$

The LASSO: a constrained optimization problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} \|\mathbf{w}^\top \mathbf{x}_i - y_i\|^2 \\ \text{subject to} & \|\mathbf{w}\|_1 \le R, \end{array}$$

here we minimize the w <=R
the weight matrix is bounded.



where $R \in \mathbb{R}_+$ is some parameter to control the bias-variance tradeoff (R large: low bias, high variance; R small: large bias, low variance)

properties of R

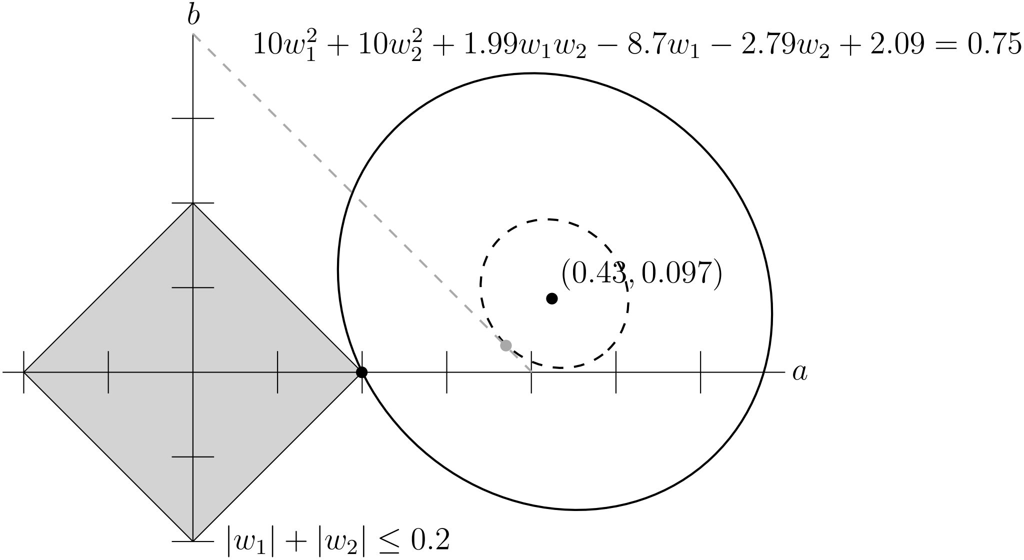
$\|\mathbf{w}\|_1 = \sum_{i=1}^{d} |w_j|$ is the 1-norm.

In our case:

$$\begin{array}{ll} \text{minimize} & f(w_1, w_2) = 10w_1^2 + 10w_2^2 + 1.99w_1w_2 - 8.7w_1 - 2.79w_2 + 2.09 \\ \text{subject to} & |w_1| + |w_2| \le R, \end{array}$$

$R = 0.2 \Rightarrow \mathbf{w}^\star = (w_1^\star, w_2^\star) = (0.2, 0)$: TOEFL is gone! But large bias 0.2 vs. 0.43
$R = 0.3 \Rightarrow \mathbf{w}^\star = (w_1^\star, w_2^\star) = (0.3, 0)$: TOEFL is still gone and bias better!
$R = 0.4 \Rightarrow \mathbf{w}^\star = (w_1^\star, w_2^\star) = (0.36, 0.036)$: TOEFL creeps back in
$R \ge 0.6 \Rightarrow \mathbf{w}^\star = (w_1^\star, w_2^\star) = (0.43, 0.097)$: original least squares solution

Geometry of the LASSO

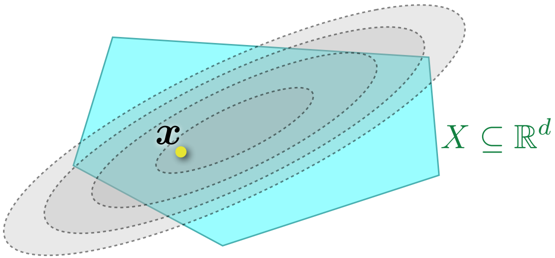


Constrained Optimization

Solving Constrained Optimization Problems

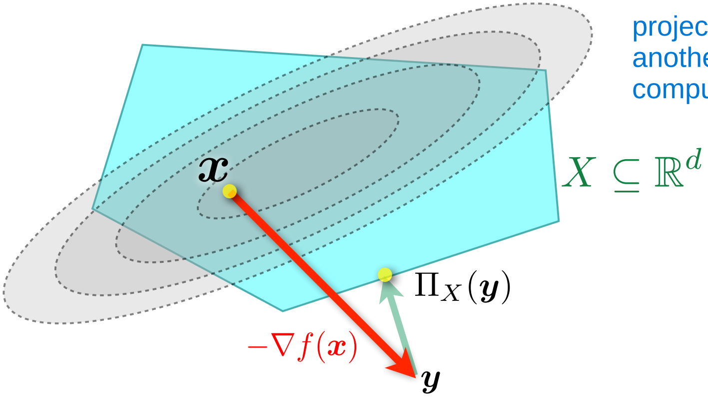
$$\begin{array}{rl} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in X \end{array}$$

- Here: Projected Gradient Descent



Projected Gradient Descent

Idea: project onto X after every step: $\Pi_X(\mathbf{y}) := \operatorname{argmin}_{\mathbf{x}\in X} \|\mathbf{x}-\mathbf{y}\|$



projecting the onto another space for better computation.

Projected gradient descent: $\mathbf{x}_{t+1} := \Pi_X\left[\mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t)\right]$

The Algorithm

Projected gradient descent: choose $\mathbf{x}_0 \in \mathbb{R}^d$.

letting the original gradient be the intermediate variable, used for computing the X_{t+1}

$$\begin{aligned}\mathbf{y}_{t+1} &:= \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t),\\ \mathbf{x}_{t+1} &:= \Pi_X(\mathbf{y}_{t+1}) := \operatorname*{argmin}_{\mathbf{x}\in X} \|\mathbf{x}-\mathbf{y}_{t+1}\|^2\end{aligned}$$

this is the projecting function.

for **times** $t = 0, 1, \ldots,$ and **stepsize** $\gamma \geq 0$.

Properties of Projection

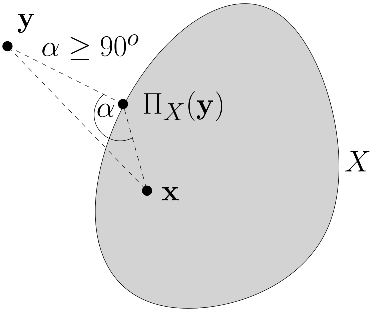
Fact 4.1

Let $X \subseteq \mathbb{R}^d$ be closed and convex, $\mathbf{x} \in X, \mathbf{y} \in \mathbb{R}^d$. Then

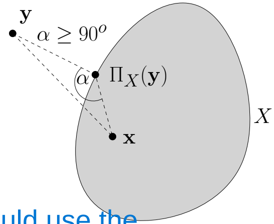
(i) $(\mathbf{x} - \Pi_X(\mathbf{y}))^\top (\mathbf{y} - \Pi_X(\mathbf{y})) \leq 0$.

(ii) $\|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2$.

these are useful properties for the exam questions.



Properties of Projection II



Fact 4.1

Let $X \subseteq \mathbb{R}^d$ be closed and convex, $\mathbf{x} \in X, \mathbf{y} \in \mathbb{R}^d$. Then

(i) $(\mathbf{x} - \Pi_X(\mathbf{y}))^\top(\mathbf{y} - \Pi_X(\mathbf{y})) \leq 0$.

(ii) $\|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2$.

this means that this function could use the properties of convex functions.

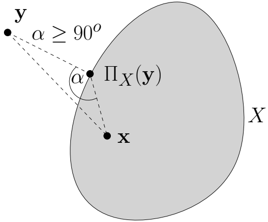
Proof.

(i) $\Pi_X(\mathbf{y})$ is minimizer of (differentiable) convex function $d_{\mathbf{y}}(\mathbf{x}) = \|\mathbf{x} - \mathbf{y}\|^2$ over X.
By first-order characterization of optimality (Lemma 2.28),

$$
\begin{array}{rcl}
0 & \leq & \nabla d_{\mathbf{y}}(\Pi_X(\mathbf{y}))^\top(\mathbf{x} - \Pi_X(\mathbf{y})) \\
& = & 2(\Pi_X(\mathbf{y}) - \mathbf{y})^\top(\mathbf{x} - \Pi_X(\mathbf{y})) \\
\Leftrightarrow \quad 0 & \geq & 2(\mathbf{y} - \Pi_X(\mathbf{y}))^\top(\mathbf{x} - \Pi_X(\mathbf{y})) \\
\Leftrightarrow \quad 0 & \geq & (\mathbf{x} - \Pi_X(\mathbf{y}))^\top(\mathbf{y} - \Pi_X(\mathbf{y}))
\end{array}
$$

$\square$

Properties of Projection III



Fact 4.1

Let $X \subseteq \mathbb{R}^d$ be closed and convex, $\mathbf{x} \in X, \mathbf{y} \in \mathbb{R}^d$. Then

(i) $(\mathbf{x} - \Pi_X(\mathbf{y}))^\top (\mathbf{y} - \Pi_X(\mathbf{y})) \leq 0$.

(ii) $\|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2$.

Proof.

(ii)

$$\mathbf{v} := (\mathbf{x} - \Pi_X(\mathbf{y})), \quad \mathbf{w} := (\mathbf{y} - \Pi_X(\mathbf{y})).$$

By (i),

$$\begin{aligned} 0 \geq 2\mathbf{v}^\top \mathbf{w} &= \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2 \\ &= \|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 - \|\mathbf{x} - \mathbf{y}\|^2. \end{aligned}$$

□

Results for projected gradient descent over closed and convex X

The same number of steps as gradient descent over $\mathbb{R}^d$!

it has the same number of steps with GD.

- Lipschitz convex functions over X: $\mathcal{O}(1/\varepsilon^2)$ steps
- Smooth convex functions over X: $\mathcal{O}(1/\varepsilon)$ steps
- Smooth and strongly convex functions over X: $\mathcal{O}(\log(1/\varepsilon))$ steps

We will adapt the previous proofs for gradient descent.

BUT:

this takes a longer time due to the projection.

- Each step involves a projection onto X
- This may or may not be efficient (in relevant cases, it is)...

Here: Analysis for smooth convex functions over X.

For the other cases, see the lecture notes.

Projected Sufficient decrease

Recall: f is smooth (with parameter L) over X if

this is the theorithm.

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X.$$

Lemma 4.3

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be differentiable and smooth with parameter L over a closed and convex set $X \subseteq \mathbf{dom}(f)$. Choosing stepsize

$$\gamma := \frac{1}{L},$$

projected gradient descent with arbitrary $\mathbf{x}_0 \in X$ satisfies

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2, \quad t \geq 0.$$

Projected Sufficient decrease II

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2.$$

Proof.

Use smoothness, $\mathbf{y}_{t+1} - \mathbf{x}_t = -\nabla f(\mathbf{x}_t)/L$, $2\mathbf{v}^\top\mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2$:

$$\begin{aligned}
f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\
&= f(\mathbf{x}_t) - L(\mathbf{y}_{t+1} - \mathbf{x}_t)^\top(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\
&= f(\mathbf{x}_t) - \frac{L}{2}\left(\|\mathbf{y}_{t+1} - \mathbf{x}_t\|^2 + \underline{\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2} - \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2\right) + \frac{L}{2}\underline{\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2} \\
&= f(\mathbf{x}_t) - \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_t\|^2 + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2 \\
&= f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2.
\end{aligned}$$

Smooth convex functions over X: $\mathcal{O}(1/\varepsilon)$ steps

Theorem 4.4

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be convex and differentiable. Let $X \subseteq \mathbf{dom}(f)$ be a closed convex set, and assume that there is a minimizer $\mathbf{x}^\star$ of f over X; furthermore, suppose that f is smooth over X with parameter L. Choosing stepsize

$$\gamma := \frac{1}{L},$$

projected gradient descent yields

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

Exactly the same bound as in the unconstrained case!

Smooth convex functions over X: $\mathcal{O}(1/\varepsilon)$ steps II

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

Before, we used sufficient decrease to bound sum of squared gradients in the vanilla analysis:

$$\frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 \leq f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2$$

But now: projected sufficient decrease has an extra term $\frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2$.

Compensate in the vanilla analysis for this!

Constrained vanilla analysis

- Replace $\mathbf{x}_{t+1}$ in the vanilla analysis with $\mathbf{y}_{t+1}$ (the unprojected gradient step):

$$\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) = \frac{1}{2\gamma}\left(\gamma^2\|\mathbf{g}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{y}_{t+1} - \mathbf{x}^\star\|^2\right).$$

- Use Fact 4.1 (ii): $\quad \|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 \le \|\mathbf{x} - \mathbf{y}\|^2.$
- With $\mathbf{x} = \mathbf{x}^\star, \mathbf{y} = \mathbf{y}_{t+1}$, we have $\Pi_X(\mathbf{y}) = \mathbf{x}_{t+1}$, and hence

$$\|\mathbf{x}^\star - \mathbf{x}_{t+1}\|^2 + \underline{\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2} \quad \le \|\mathbf{x}^\star - \mathbf{y}_{t+1}\|^2$$

- We get back to the standard vanilla analysis... but with a saving!

$$\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) \le \frac{1}{2\gamma}\left(\gamma^2\|\mathbf{g}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 - \underline{\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2}\right)$$

Smooth convex functions over X: $\mathcal{O}(1/\varepsilon)$ steps III

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

Proof.

Use $f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)$ (convexity), vanilla analysis with saving, $\gamma = 1/L$:

$$\begin{aligned}
\sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) &\leq \sum_{t=0}^{T-1}\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) \\
&\leq \frac{1}{2L}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2 + \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 - \underline{\frac{L}{2}\sum_{t=0}^{T-1}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2}.
\end{aligned}$$

Use projected sufficient decrease to bound $\frac{1}{2L}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2$ by

$$\sum_{t=0}^{T-1}\left(f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) + \frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2\right) = f(\mathbf{x}_0) - f(\mathbf{x}_T) + \underline{\frac{L}{2}\sum_{t=0}^{T-1}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2}.$$

Smooth convex functions over X: $\mathcal{O}(1/\varepsilon)$ steps IV

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

Proof.

Putting it together: extra terms cancel, and as in unconstrained case, we get

$$\sum_{t=1}^{T}\left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) \leq \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

Exercise 32: again, we make progress in every step (not immediate from projected sufficient decrease). Hence,

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{1}{T}\sum_{t=1}^{T}\left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

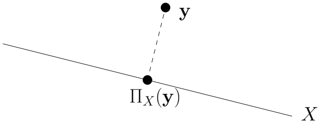
□

The Projection Step: $\Pi_X(\mathbf{y}) := \mathrm{argmin}_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{y}\|$

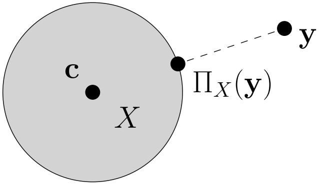
Computing $\Pi_X(\mathbf{y})$ is an optimization problem itself.

It can efficiently be solved in relevant cases:

- Projecting onto an affine subspace (leads to system of linear equations, similar to least squares)

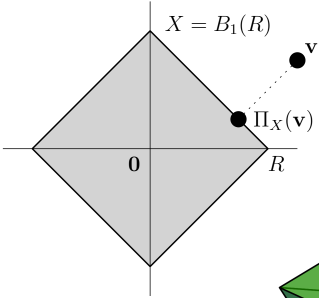


- Projecting onto a Euclidean ball with center $\mathbf{c}$ (simply scale the vector $\mathbf{y} - \mathbf{c}$)

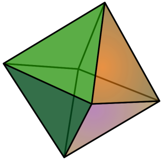


Projecting onto ℓ_1-balls (needed in LASSO)

W.l.o.g. restrict to center at $\mathbf{0}$: $B_1(R) = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_1 = \sum_{i=1}^d |x_i| \leq R\}$.



$B_1(R)$ is the cross polytope ($2d$ vertices, 2^d facets).



(octahedron, $d = 3$)

Section 4.5: projection can be computed in $\mathcal{O}(d \log d)$ time (can be improved to $\mathcal{O}(d)$)

Bibliography

Judith Zimmermann.
Information Processing for Effective and Stable Admission.
PhD thesis, ETH Zurich, 2016.