Learning objectives. By the end of class, you should be able to

1. derive the projection onto the nonnegative orthant by deciding which walls are active and solving the KKT conditions;
2. derive the descent lemma for projected gradient descent,
3. prove that projGD converges to a KKT point at the rate of gradient descent.

1 Recap: How to deal with constraints

Consider maximizing $x_1 + x_2$ over the unit disc,

$$\min_{x \in \mathbb{R}^2} \; -x_1 - x_2 \text{ s.t. } x_1^2 + x_2^2 \leq 1.$$

Penalty and barrier methods turn it into an unconstrained problem.

- Quadratic penalty, with $(t)_+ = \max\{t, 0\}$,

$$F_\mu(x) = -x_1 - x_2 + \frac{1}{2\mu}\left[(x_1^2 + x_2^2 - 1)_+\right]^2.$$

- Logarithmic barrier,

$$F_\mu(x) = -x_1 - x_2 - \mu \log(1 - x_1^2 - x_2^2).$$

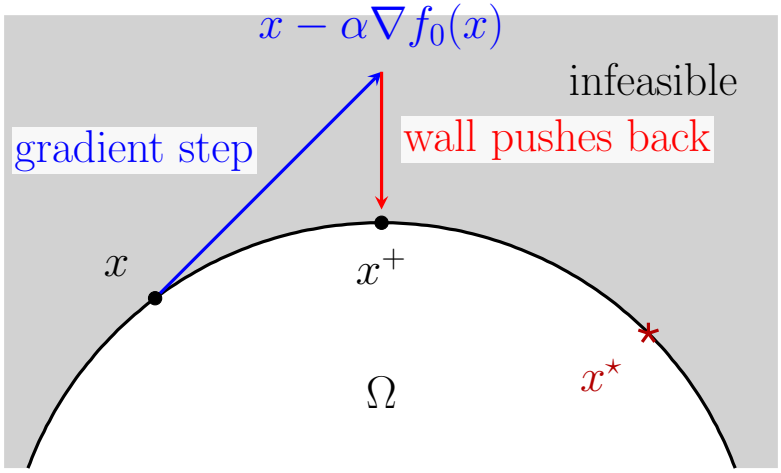
For every $\mu > 0$, the penalty minimizer lies outside the disc, and the barrier minimizer lies strictly inside. Both need $\mu \to 0$.

The disc Ω has a cheap **projection**, its closest point,

$$\mathrm{proj}_\Omega(u) = \arg\min_{x\in\Omega} \frac{1}{2}\|x-u\|^2 = \begin{cases} u, & \|u\| \le 1, \\ u/\|u\|, & \|u\| > 1. \end{cases}$$

Projected gradient descent with step size $\alpha > 0$ takes a gradient step and projects it back,

$$x^+ = \mathrm{proj}_\Omega\left(x - \alpha\nabla f_0(x)\right) = \mathrm{proj}_\Omega\left(x + \alpha\begin{bmatrix}1\\1\end{bmatrix}\right).$$



- From $x = (-0.6, 0.8)$ with $\alpha = 0.6$, the gradient step lands at $(0, 1.4)$, outside the disc. The projection returns it to $x^+ = (0, 1)$.
- Repeating, the iterates move along the circle toward $x^\star = (1/\sqrt{2}, 1/\sqrt{2})$.

Question 1. Why could projected gradient descent be better than penalty and barrier methods?

2 Projections by KKT

Assume the set below is **nonempty** and satisfies **Slater's condition**

$$\Omega = \{x \in \mathbb{R}^n : f_i(x) \leq 0,\ h_j(x) = 0 \text{ for all } i, j\},$$

with $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, p\}$.

Question 2. What does that mean for the individual functions f_i and h_j?

The projection of $u \in \mathbb{R}^n$ is a constrained optimization problem,

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \frac{1}{2}\|x - u\|^2 \\ \text{s.t.} \quad & f_i(x) \leq 0 \text{ for all } i \in \{1, \ldots, m\}, \\ & h_j(x) = 0 \text{ for all } j \in \{1, \ldots, p\}. \end{aligned}$$

The projection problem is convex and satisfies Slater's condition. Strong duality holds, and the KKT conditions are necessary and sufficient for optimality. The projection is unique, and every solution of the KKT conditions yields this projection, although the Lagrange multipliers need not be unique.

Question 3. Write out the KKT conditions of the projection problem.

Question 4. What happens when u is already in Ω?

To find a projection, use complementarity. Each wall is either

$$\begin{aligned} \text{inactive:} \quad & f_i(x) < 0 \quad \text{and} \quad y_i = 0, \\ \text{active:} \quad & f_i(x) = 0 \quad \text{and} \quad y_i \geq 0. \end{aligned}$$

The task amounts to checking which wall is active.

Lemma 1. Orthant. Let $\Omega = \{x \in \mathbb{R}^n : x_i \geq 0 \text{ for all } i\}$. Then

$$[\text{proj}_\Omega(u)]_i = \max\{u_i, 0\} = \begin{cases} u_i, & u_i \geq 0, \\ 0, & u_i \leq 0. \end{cases}$$

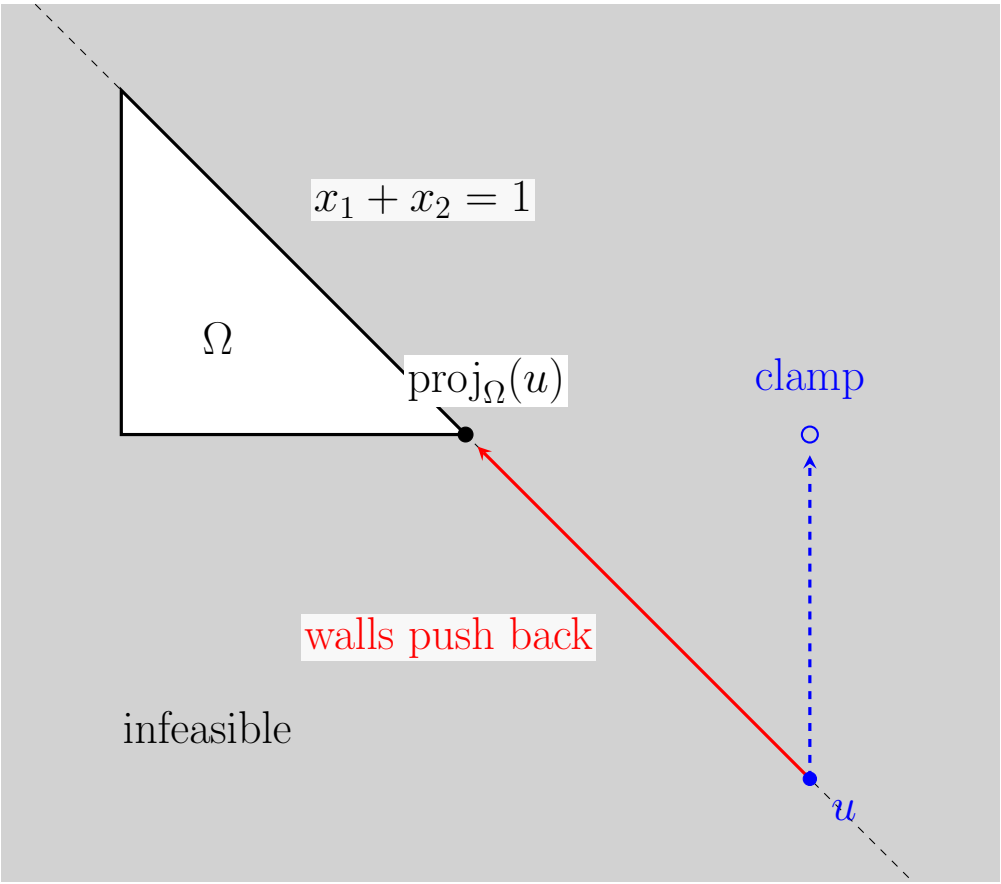
Example 2. Projection onto a triangle. Let

$$\Omega = \left\{ x \in \mathbb{R}^2 : f_1(x) \le 0,\ f_2(x) \le 0,\ f_3(x) \le 0 \right\},$$

with walls

$$f_1(x) = -x_1, \qquad f_2(x) = -x_2, \qquad f_3(x) = x_1 + x_2 - 1.$$

Find $\mathrm{proj}_\Omega(u)$ for $u = (2, -1)$.



- The point $u = (2, -1)$ lies below the floor $x_2 = 0$, on the line $x_1 + x_2 = 1$.
- Clamping onto the orthant alone gives $(2, 0)$, still outside Ω. The closest point of Ω is the corner $(1, 0)$.

Without a good guess, we must check the KKT conditions for every set of active walls. Active walls hold with equality, inactive walls have $y_i = 0$, and stationarity fixes x,

active walls	x	fails
none	$(2, -1)$	$f_2(x) = 1 > 0$
$\{1\}$	$(0, -1)$	$f_2(x) = 1 > 0$
$\{2\}$	$(2, 0)$	$f_3(x) = 1 > 0$
$\{3\}$	$(2, -1)$	$f_2(x) = 1 > 0$
$\{1, 2\}$	$(0, 0)$	$y_1 = -2 < 0$
$\{1, 3\}$	$(0, 1)$	$y_1 = -4 < 0$
$\{2, 3\}$	$(1, 0)$	nothing
$\{1, 2, 3\}$	no x	

With m walls that share variables, there are 2^m sets to check.

Question 5. Why don't we use projected gradient descent everywhere?

3 Convergence of projected gradient descent

Let f have L_1-Lipschitz gradient. Unconstrained GD works by minimizing the quadratic upper bound,

$$\begin{aligned}\arg\min_{x^+\in\mathbb{R}^n}\left\{f(x)+\nabla f(x)^T(x^+-x)+\frac{L_1}{2}\left\|x-x^+\right\|^2\right\}\\ =x-\frac{1}{L_1}\nabla f(x)\end{aligned}$$

and the associated decrement is the **descent lemma**

$$f(x^+)\le f(x)-\frac{1}{2L_1}\left\|\nabla f(x)\right\|^2.$$

Now, let $\Omega=\{x: f_i(x)\le 0,\ h_j(x)=0\}$ be nonempty and satisfy Slater's condition. Minimizing the same upper-bound over the constrained set yields

$$\begin{aligned}\arg\min_{x^+\in\Omega}\left\{f_0(x)+\nabla f_0(x)^T(x^+-x)+\frac{L_1}{2}\left\|x-x^+\right\|^2\right\}\\ =\mathrm{proj}_\Omega\left(x-\frac{1}{L_1}\nabla f_0(x)\right).\end{aligned}$$

Indeed, completing the square yields

$$\begin{aligned}&f_0(x)+\nabla f_0(x)^T(x^+-x)+\frac{L_1}{2}\left\|x^+-x\right\|^2\\ &=\frac{L_1}{2}\left\|x^+-\left(x-\frac{1}{L_1}\nabla f_0(x)\right)\right\|^2+\underbrace{f_0(x)-\frac{1}{2L_1}\left\|\nabla f_0(x)\right\|^2}_{\text{constant}}.\end{aligned}$$

Define the **gradient map**

$$g=L_1(x-x^+),\qquad \text{so that}\qquad x^+=x-\frac{1}{L_1}g.$$

We will show that ProjGD satisfies the ***projected*** **descent lemma**

$$f(x^+) \le f(x) - \frac{1}{2L_1} \|g\|^2 .$$

From which essentially all conclusions of unconstrained GD extend to ProjGD.

Lemma 3. KKT conditions of the step. Let $x^+ = \text{proj}_\Omega(x - \frac{1}{L_1}\nabla f_0(x))$ and $g = L_1(x - x^+)$. Then there are multipliers $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^p$ with

$$\text{(primal feasibility) } f_i(x^+) \le 0, \quad h_j(x^+) = 0 \text{ for all } i, j,$$

$$\text{(stationarity) } \nabla f_0(x) + \sum_{i=1}^m y_i \nabla f_i(x^+) + \sum_{j=1}^p z_j \nabla h_j(x^+) = g,$$

$$\text{(dual feasibility) } y_i \ge 0 \text{ for all } i,$$

$$\text{(complementarity) } y_i f_i(x^+) = 0 \text{ for all } i.$$

Lemma 4. Stationarity. Let (y, z) be the multipliers of the step. Then

$$\left\|\nabla_x \mathscr{L}(x^+, y, z)\right\| \le 2 \|g\| .$$

The descent lemma generalizes to the projected descent lemma.

Lemma 5. Projected descent lemma. Let $x \in \Omega$. Then

$$f_0(x^+) \leq f_0(x) - \frac{1}{2L_1} \|g\|^2 .$$

Theorem 6. KKT convergence. Let f_0 have L_1-Lipschitz gradient. Then projected gradient descent

$$x_{k+1} = \text{proj}_\Omega\left(x_k - \frac{1}{L_1}\nabla f_0(x_k)\right)$$

from $x_0 \in \Omega$ attains an iterate $x_{\text{best}} \in \{x_1, \ldots, x_N\}$ and multipliers $y \in \mathbb{R}^m$, $z \in \mathbb{R}^p$ satisfying

$$\text{(primal feasibility) } f_i(x_{\text{best}}) \le 0, \quad h_j(x_{\text{best}}) = 0 \text{ for all } i, j,$$

$$\text{(stationarity) } \|\nabla_x \mathscr{L}(x_{\text{best}}, y, z)\| \le 2\sqrt{\frac{2L_1\left[f_0(x_0) - p^\star\right]}{N}},$$

$$\text{(dual feasibility) } y_i \ge 0 \text{ for all } i,$$

$$\text{(complementarity) } y_i f_i(x_{\text{best}}) = 0 \text{ for all } i.$$

Question 6. Does the theorem require f_0 to be convex?

Question 7. Why is x_{best} taken from $\{x_1, \ldots, x_N\}$, and not from $\{x_0, \ldots, x_{N-1}\}$ as in gradient descent?

Compare the two guarantees after N steps,

$$\text{gradient descent:} \quad \|\nabla f(x_{\text{best}})\| \leq \sqrt{\frac{2L_1\left[f(x_0) - f_{\text{inf}}\right]}{N}},$$

$$\text{projected:} \quad \|\nabla_x \mathscr{L}(x_{\text{best}}, y, z)\| \leq 2\sqrt{\frac{2L_1\left[f_0(x_0) - p^\star\right]}{N}}.$$

Same rate, same proof, with the gradient map in place of the gradient.

4 Example: Nonnegative least squares

Nonnegative least squares fits $b \approx Ax$ with $x \geq 0$. For example, b is the measured spectrum of a mixture, the columns of A are the spectra of known substances, and x holds their concentrations,

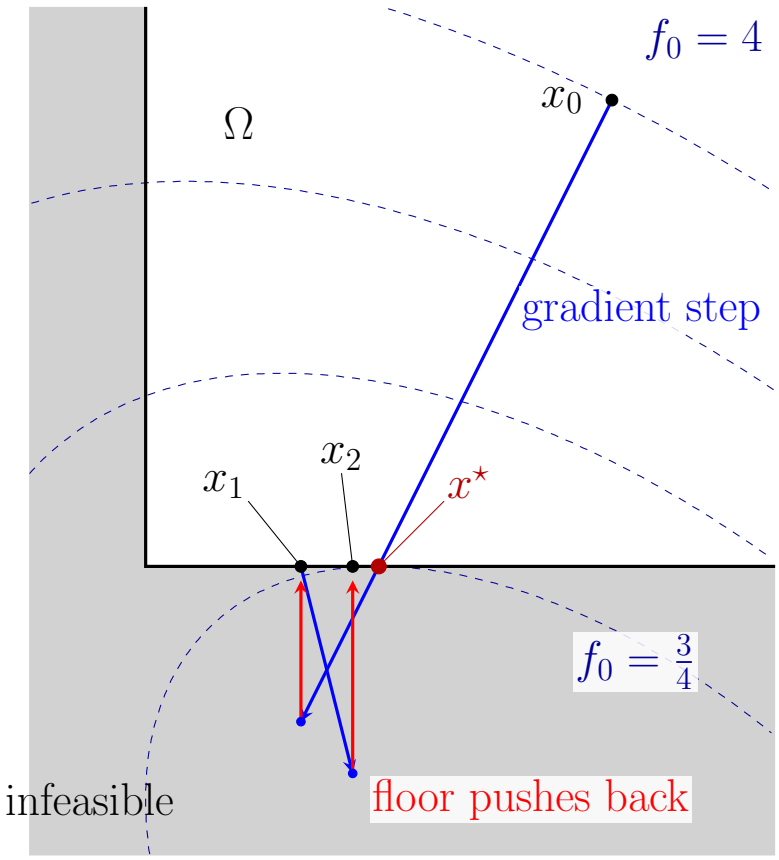
$$\min_x \ \frac{1}{2}\|Ax - b\|^2 \text{ s.t. } x \geq 0.$$

The gradient is $A^T(Ax - b)$, and projecting onto $x \geq 0$ clamps each coordinate at 0. Projected gradient descent reads

$$x_{k+1} = \max\left\{x_k - \frac{1}{L_1}A^T(Ax_k - b),\, 0\right\}, \qquad L_1 = \lambda_{\max}(A^TA),$$

with the maximum taken coordinate by coordinate. The figure runs it for

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \qquad L_1 = 3, \qquad x_0 = (1, 1).$$



- The dashed curves are level sets of f_0, centered at the unconstrained minimizer $(1, -1)$, outside Ω.
- Each gradient step goes below the floor $x_2 = 0$, and the floor pushes it back up.
- The constrained minimizer $x^\star = (\frac{1}{2}, 0)$ is where the level set $f_0 = \frac{3}{4}$ touches the floor.

Question 8. What does projected gradient descent do once the iterates are on the floor $x_2 = 0$?

5 Summary

Projections. Under Slater's condition, the projection onto Ω is the solution of the KKT conditions of

$$\min_x \frac{1}{2} \|x - u\|^2 \text{ s.t. } x \in \Omega.$$

For the orthant, the walls decouple, and the projection clamps each coordinate at 0. When walls share variables, the active walls must be guessed together, and the projection is as hard as a constrained problem.

Projected gradient descent. With the multipliers of the step, every iterate satisfies primal feasibility, dual feasibility, and complementarity exactly. The gradient map $g = L_1(x - x^+)$ takes the place of the gradient, and only stationarity is approximate,

$$f_0(x^+) \leq f_0(x) - \frac{1}{2L_1} \|g\|^2, \qquad \left\|\nabla_x \mathscr{L}(x^+, y, z)\right\| \leq 2 \|g\|.$$

Summing the decrease gives the rate of gradient descent,

$$\|\nabla_x \mathscr{L}(x_{\text{best}}, y, z)\| \leq 2\sqrt{\frac{2L_1 [f_0(x_0) - p^\star]}{N}}.$$