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,

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

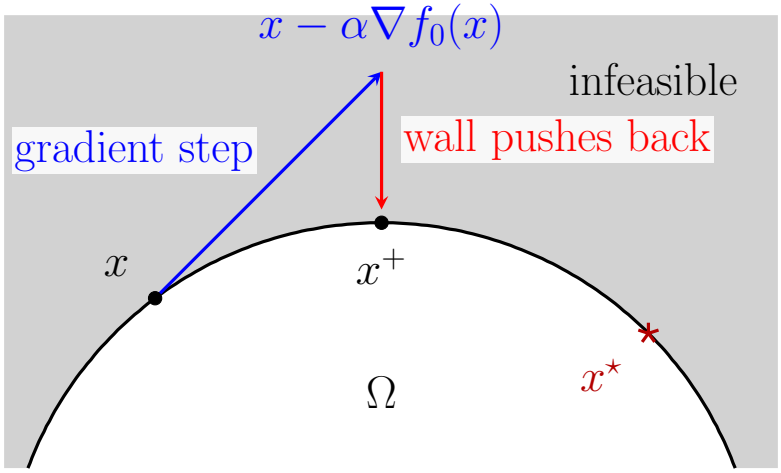
For every $\mu > 0$, the penalty minimizer lies outside the disc, and the barrier minimizer lies strictly inside. Both need $\mu \to 0$.

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?

Solution: Every iterate is feasible, $x_k \in \Omega$ exactly. There is no μ to drive to 0.

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?

Solution: Every f_i is convex, and every h_j is affine. Some $\bar{x} \in \Omega$ has $f_i(\bar{x}) < 0$ for every f_i that is not affine.

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.

Solution: The gradient of $\frac{1}{2}\|x-u\|^2$ is $x-u$.

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

$$\text{(stationarity) } x - u + \sum_{i=1}^{m} y_i \nabla f_i(x) + \sum_{j=1}^{p} z_j \nabla h_j(x) = 0,$$

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

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

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

Solution: The projection is u itself. The point $x = u$ is feasible, so we can set $y = 0$ and $z = 0$. Stationarity reads $u - u = 0$, and complementarity reads $0 \cdot f_i(u) = 0$.

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

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

Proof: The walls are $f_i(x) = -x_i \leq 0$, and the KKT conditions of the projection, coordinate by coordinate, are

$$\begin{gathered} \text{(stationarity) } x_i - u_i - y_i = 0, \\ \text{(primal feasibility) } -x_i \leq 0, \\ \text{(dual feasibility) } y_i \geq 0, \\ \text{(complementarity) } y_i \cdot (-x_i) = 0. \end{gathered}$$

Coordinate i involves only x_i and y_i, so each wall can be checked on its own.

Wall i inactive. $x_i > 0$ and $y_i = 0$. Stationarity gives

$$x_i - u_i - 0 = 0 \quad \Longrightarrow \quad x_i = u_i.$$

Consistent with $x_i > 0$ when $u_i > 0$.

Wall i active. $x_i = 0$ and $y_i \geq 0$. Stationarity gives

$$0 - u_i - y_i = 0 \quad \Longrightarrow \quad y_i = -u_i.$$

Consistent with $y_i \geq 0$ when $u_i \leq 0$. □

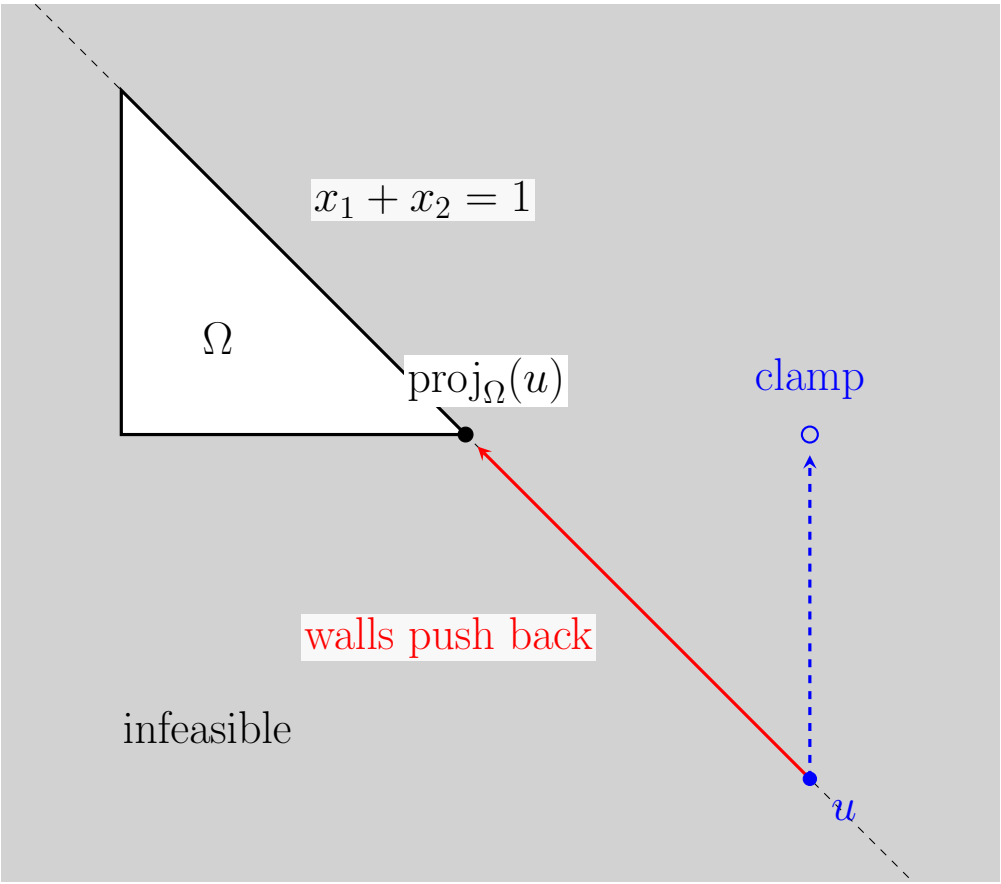
Example 2. Projection onto a triangle. Let

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

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 $(1, 0)$.

Solution: The Lagrangian of the projection is

$$\mathscr{L}(x,y) = \frac{1}{2}\left[(x_1-2)^2+(x_2+1)^2\right] - y_1x_1 - y_2x_2 + y_3(x_1+x_2-1).$$

Its gradient in x, coordinate by coordinate,

$$\frac{\partial \mathscr{L}}{\partial x_1} = x_1 - 2 - y_1 + y_3, \qquad \frac{\partial \mathscr{L}}{\partial x_2} = x_2 + 1 - y_2 + y_3.$$

The KKT conditions,

$$\begin{gathered}
\text{(stationarity) } x_1 - 2 - y_1 + y_3 = 0, \qquad x_2 + 1 - y_2 + y_3 = 0, \\
\text{(primal feasibility) } -x_1 \le 0, \quad -x_2 \le 0, \quad x_1 + x_2 - 1 \le 0, \\
\text{(dual feasibility) } y_1 \ge 0, \quad y_2 \ge 0, \quad y_3 \ge 0, \\
\text{(complementarity) } y_1x_1 = 0, \quad y_2x_2 = 0, \quad y_3(x_1+x_2-1) = 0.
\end{gathered}$$

From the figure, guess $x = (1,0)$. Primal feasibility holds,

$$f_1(x) = -1 < 0, \qquad f_2(x) = 0, \qquad f_3(x) = 1 + 0 - 1 = 0,$$

so wall 1 is inactive, and walls 2 and 3 are active. Complementarity gives $y_1 = 0$. Stationarity gives the pushes of the active walls,

$$\begin{aligned}
1 - 2 - 0 + y_3 = 0 \quad &\Longrightarrow \quad y_3 = 1, \\
0 + 1 - y_2 + 1 = 0 \quad &\Longrightarrow \quad y_2 = 2.
\end{aligned}$$

Dual feasibility holds, $y = (0,2,1) \ge 0$. All four KKT conditions hold, so $\text{proj}_\Omega(2,-1) = (1,0)$.

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?

Solution: Every step is a projection, a constrained problem with the same constraints. Unless the walls decouple, as for the orthant, it has no formula. Even $\text{proj}_\Omega(0)$ finds a feasible point of Ω.

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 $x^+ = \mathrm{proj}_\Omega(x - \frac{1}{L_1}\nabla f_0(x))$ and $g = L_1(x - x^+)$. Then there are multipliers $y \in \mathbb{R}^m$ and $z \in \mathbb{R}^p$ with

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

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

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

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

Proof: The KKT conditions of the projection, with multipliers written as y_i/L_1 and z_j/L_1,

$$x^+ - x + \frac{1}{L_1}\nabla f_0(x) + \sum_{i=1}^m \frac{y_i}{L_1}\nabla f_i(x^+) + \sum_{j=1}^p \frac{z_j}{L_1}\nabla h_j(x^+) = 0.$$

Multiply by L_1, and substitute $L_1(x - x^+) = g$. □

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

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

Proof: Stationarity of the step replaces the two sums,

$$\nabla_x \mathscr{L}(x^+, y, z) = \nabla f_0(x^+) - \nabla f_0(x) + g.$$

Lipschitz gradient, with $L_1 \|x^+ - x\| = \|g\|$,

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

□

The descent lemma generalizes to the projected descent lemma.

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

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

Proof: Inner product with g. Solve stationarity of the step for $\nabla f_0(x)$,

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

Take the inner product with g,

$$\nabla f_0(x)^T g = \|g\|^2 - \sum_{i=1}^{m} y_i \nabla f_i(x^+)^T g - \sum_{j=1}^{p} z_j \nabla h_j(x^+)^T g.$$

Inequality walls. Each f_i is convex, so it lies above its tangent at x^+,

$$f_i(x) \geq f_i(x^+) + \nabla f_i(x^+)^T (x - x^+).$$

Multiply by $L_1 > 0$, and use $g = L_1(x - x^+)$,

$$\nabla f_i(x^+)^T g \leq L_1 \left[f_i(x) - f_i(x^+) \right].$$

Multiply by $y_i \geq 0$, then use $y_i f_i(x^+) = 0$ and $f_i(x) \leq 0$,

$$y_i \nabla f_i(x^+)^T g \leq L_1 y_i f_i(x) - L_1 \underbrace{y_i f_i(x^+)}_{=0} = L_1 \underbrace{y_i}_{\geq 0} \underbrace{f_i(x)}_{\leq 0} \leq 0.$$

Equality walls. Each h_j is affine, so it equals its tangent, and $h_j(x) = h_j(x^+) = 0$,

$$\nabla h_j(x^+)^T g = L_1 \nabla h_j(x^+)^T (x - x^+) = L_1 \left[h_j(x) - h_j(x^+) \right] = 0.$$

Combine. Every wall term is ≤ 0 or $= 0$, so

$$\nabla f_0(x)^T g \geq \|g\|^2 .$$

Upper bound. Substitute $x^+ - x = -\frac{1}{L_1} g$ into the quadratic upper bound,

$$\begin{aligned} f_0(x^+) &\leq f_0(x) + \nabla f_0(x)^T (x^+ - x) + \frac{L_1}{2} \left\| x^+ - x \right\|^2 \\ &= f_0(x) - \frac{1}{L_1} \nabla f_0(x)^T g + \frac{L_1}{2} \cdot \frac{1}{L_1^2} \|g\|^2 \\ &\leq f_0(x) - \frac{1}{L_1} \|g\|^2 + \frac{1}{2L_1} \|g\|^2 \\ &= f_0(x) - \frac{1}{2L_1} \|g\|^2 . \end{aligned}$$

□

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

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

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

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

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

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

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

Proof: Write $g_k = L_1(x_k - x_{k+1})$. Sum the projected descent lemma, as for gradient descent,

$$\frac{1}{2L_1}\sum_{k=0}^{N-1}\|g_k\|^2 \le f_0(x_0) - f_0(x_N) \le f_0(x_0) - p^\star.$$

The smallest gradient map, at $k = j$, is no larger than the average,

$$\|g_j\|^2 \le \frac{2L_1\left[f_0(x_0) - p^\star\right]}{N}.$$

Take $x_{\mathrm{best}} = x_{j+1}$ with the multipliers of the step from x_j,

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

□

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

Solution: No. Only Ω must be convex. For nonconvex f_0, the KKT point need not be a global minimizer.

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?

Solution: The KKT conditions hold after a projection. Every x_{k+1} is a projection, but x_0 has not been projected yet.

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,

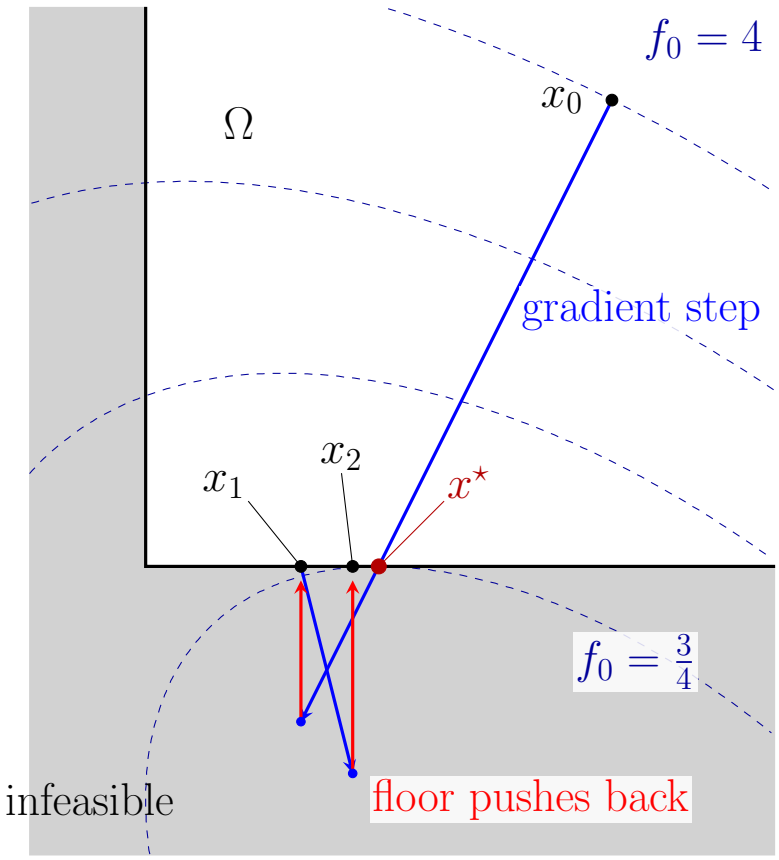
$$\min_x \ \frac{1}{2}\|Ax - b\|^2 \text{ s.t. } x \geq 0.$$

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 $x_2 = 0$?

Solution: Every gradient step points below the floor and is clamped back, so x_2 stays 0. Only x_1 moves. It is gradient descent along the floor, toward $x^\star$.

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,

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