Statistical Optimization: Lecture 12

Penalty and Augmented Lagrangian Methods

Zijian Guo

Zhejiang University
Center for Data Science

May 2, 2026

Objective

Goal. We want to solve constrained optimization problems.

Main idea. We will compare three viewpoints:

- dual updates for enforcing feasibility,
- penalty terms for discouraging constraint violation,
- the augmented Lagrangian method, which combines both.

Route of this lecture.

- Equality-constrained and general constrained problems
- Dual-ascent intuition
- Quadratic penalty method
- Augmented Lagrangian method

Problem setup

$$
\begin{aligned}
\min_{\theta} \quad & f_0(\theta) \\
\text{s.t.} \quad & h_i(\theta) = 0, \qquad i \in \mathcal{E}.
\end{aligned}
\tag{1}
$$

This is the equality-constrained problem.

$$
\begin{aligned}
\min_{\theta} \quad & f_0(\theta) \\
\text{s.t.} \quad & h_i(\theta) = 0, \qquad i \in \mathcal{E}, \\
& f_i(\theta) \leq 0, \qquad i \in \mathcal{I}.
\end{aligned}
\tag{2}
$$

This is the general constrained problem.

Outline

Dual-ascent intuition

Quadratic penalty method

Augmented Lagrangian method

Ordinary Lagrangian and dual-ascent idea

For the equality-constrained problem (1), define the ordinary Lagrangian

$$L(\theta, \nu) := f_0(\theta) + \sum_{i \in \mathcal{E}} \nu_i h_i(\theta). \tag{3}$$

The associated dual function is

$$g(\nu) := \inf_{\theta} L(\theta, \nu), \qquad \max_{\nu} g(\nu). \tag{4}$$

If

$$\theta(\nu) \in \arg\min_{\theta} L(\theta, \nu),$$

then, under suitable conditions, Danskin's theorem gives

$$\nabla g(\nu) = h(\theta(\nu)). \tag{5}$$

Therefore, gradient ascent on the dual function suggests the update

$$\nu^{k+1} = \nu^k + \alpha_k h(\theta_k), \qquad \theta_k \approx \arg\min_{\theta} L(\theta, \nu^k), \qquad \alpha_k > 0. \tag{6}$$

Why plain dual ascent is not enough

Without strong convexity, the minimizer of $L(\theta, \nu)$ with respect to θ may fail to be unique.

Then the dual function may become less regular.

As a result:

- convergence may be very slow,
- the iterates may oscillate,
- with poor step-size choices, the method may even fail to converge.

This motivates adding a quadratic penalty term to stabilize the primal step.

Outline

Dual-ascent intuition

Quadratic penalty method

Augmented Lagrangian method

Equality-constrained: quadratic penalty method

For the equality-constrained problem (1), define the quadratic penalty function

$$Q(\theta;\mu) := f_0(\theta) + \frac{\mu}{2}\sum_{i\in\mathcal{E}} h_i(\theta)^2, \qquad \mu > 0. \tag{6}$$

Idea. As μ increases, violations of the equality constraints are penalized more severely.

General constrained: quadratic penalty method

For the general constrained problem (2), define

$$
\begin{aligned}
Q(\theta;\mu) :=\, & f_0(\theta) + \frac{\mu}{2}\sum_{i\in\mathcal{E}} h_i(\theta)^2 \\
& + \frac{\mu}{2}\sum_{i\in\mathcal{I}} \left([f_i(\theta)]_+\right)^2, \qquad \mu > 0,
\end{aligned} \tag{7}
$$

where

$$
[y]_+ := \max(y, 0).
$$

The inequality term penalizes only the violated constraints.

Remarks on quadratic penalty subproblems

- The quadratic penalty subproblem is generally an unconstrained optimization problem, but it need not be convex.
- Even if f_0, h_i, and f_i are smooth, Q may be less smooth because of the inequality penalty term.
- For the general constrained problem, the function is typically still differentiable, but the Hessian may fail to exist at points where

$$f_i(\theta) = 0,$$

because the term $\left([f_i(\theta)]_+\right)^2$ is usually not twice differentiable there.

Pseudo-code (Quadratic penalty method)

```
Input: μ_0 > 0, tolerances τ_k ↓ 0, initial point θ_0^s
for k = 0, 1, 2, ... do
    approximately minimize Q(θ; μ_k) from θ_k^s
    until ‖∇_θ Q(θ; μ_k)‖ ≤ τ_k
    set θ_k ← obtained approximate minimizer
    if ‖h(θ_k)‖ ≤ ε_feas then
        return θ_k
    end if
    choose μ_{k+1} > μ_k
    choose new starting point θ_{k+1}^s
end for
```

In convergence analysis, we restrict our attention to the equality-constrained problem (1).

Theorem (Ideal convergence behavior)

Suppose that each θ_k is the exact global minimizer of (6) *and that*

$$\mu_k \uparrow \infty.$$

Then every limit point $\theta^\star$ of the sequence $\{\theta_k\}$ is a global minimizer of the equality-constrained problem

$$\min_\theta \quad f_0(\theta) \qquad s.t. \quad h(\theta) = 0.$$

Theorem (Algorithm convergence behavior)

Suppose that the tolerances and penalty parameters satisfy

$$\tau_k \to 0, \qquad \mu_k \uparrow \infty.$$

Let $\theta^\star$ be a limit point of the sequence $\{\theta_k\}$.

- *If $\theta^\star$ is infeasible, then it is a stationary point of $\|h(\theta)\|^2$.*
- *If $\theta^\star$ is feasible and the constraint gradients $\nabla h_i(\theta^\star)$ are linearly independent, then $\theta^\star$ is a KKT point.*

  Moreover, along any infinite subsequence $\mathcal{K}$ with $\theta_k \to \theta^\star \quad (k \in \mathcal{K})$, we have

  $$\mu_k h_i(\theta_k) \to \nu_i^\star, \qquad i \in \mathcal{E},$$

  where $\nu^\star$ is the multiplier vector that satisfies the KKT conditions for the equality-constrained problem.

Intuition for the limit $\mu_k h_i(\theta_k) \to \nu^\star$

At the approximate minimizer θ_k of the quadratic penalty subproblem, we have

$$\nabla Q(\theta_k; \mu_k) \approx 0.$$

This gives

$$\nabla f_0(\theta_k) + \sum_{i \in \mathcal{E}} \mu_k h_i(\theta_k) \nabla h_i(\theta_k) \approx 0.$$

Compare this with the KKT stationarity condition at a feasible limit point $\theta^\star$:

$$\nabla f_0(\theta^\star) + \sum_{i \in \mathcal{E}} \nu_i^\star \nabla h_i(\theta^\star) = 0.$$

So the quantity

$$\mu_k h_i(\theta_k)$$

plays the role of an *approximate multiplier*. Hence, along a subsequence with $\theta_k \to \theta^\star$, it is natural to expect

$$\mu_k h_i(\theta_k) \to \nu_i^\star.$$

Main drawback of quadratic penalty

Interpretation of the convergence result.

- The sequence may be attracted to infeasible stationary points.
- Even at feasible limit points, the method only guarantees convergence to KKT points.
- The quantities $\mu_k h_i(\theta_k)$ behave like multiplier estimates.

Main drawback.

- As μ_k increases, the Hessian matrix becomes increasingly ill-conditioned.
- The quadratic penalty function is usually not exact: even for a fixed $\mu > 0$, its minimizer does not necessarily coincide with the constrained solution.

Outline

Dual-ascent intuition

Quadratic penalty method

Augmented Lagrangian method

Why dual ascent alone can be unstable

Recall the ordinary Lagrangian

$$L(\theta, \nu^k) = f_0(\theta) + \sum_{i \in \mathcal{E}} \nu_i^k h_i(\theta).$$

Dual ascent minimizes $L(\theta, \nu^k)$ with respect to θ, and then updates ν^k by the current constraint violation.

The issue is that the constraint enters only linearly. So if θ is far from feasibility, the primal step may still be poorly behaved.

In particular, the θ-subproblem can be sensitive to the current multiplier estimate ν^k, especially when ν^k is still inaccurate.

Why the quadratic penalty helps

Now add a quadratic penalty term:

$$\frac{\mu_k}{2} \sum_{i \in \mathcal{E}} h_i(\theta)^2.$$

This penalizes large constraint violation more strongly, so the primal subproblem becomes more stable.

At an approximate minimizer θ_k, we have

$$\nabla f_0(\theta_k) + \sum_{i \in \mathcal{E}} \mu_k h_i(\theta_k) \nabla h_i(\theta_k) \approx 0.$$

So the quantity

$$\mu_k h_i(\theta_k)$$

already behaves like an approximate multiplier correction.

Another view: how the augmented Lagrangian arises

Start from the equality-constrained problem

$$\min_{\theta} f_0(\theta) \qquad \text{s.t.} \quad h_i(\theta) = 0, \quad i \in \mathcal{E}.$$

A natural idea is to add a quadratic penalty term:

$$\min_{\theta} f_0(\theta) + \frac{\mu}{2} \sum_{i \in \mathcal{E}} h_i(\theta)^2 \qquad \text{s.t.} \quad h_i(\theta) = 0, \quad i \in \mathcal{E}.$$

Now take the ordinary Lagrangian of this penalized constrained problem:

$$L_A(\theta, \nu; \mu) := f_0(\theta) + \sum_{i \in \mathcal{E}} \nu_i h_i(\theta) + \frac{\mu}{2} \sum_{i \in \mathcal{E}} h_i(\theta)^2.$$

This is exactly the augmented Lagrangian.

So ALM can also be viewed as:

$$\text{ordinary Lagrangian} \; + \; \text{quadratic penalty}.$$

Multiplier update

At iteration k, fix $\mu_k > 0$ and the current multiplier ν^k, and approximately minimize

$$L_A(\theta, \nu^k; \mu_k) \quad \text{with respect to } \theta.$$

If θ_k is an approximate minimizer, then

$$0 \approx \nabla_\theta L_A(\theta_k, \nu^k; \mu_k) = \nabla f_0(\theta_k) + \sum_{i \in \mathcal{E}} \big(\nu_i^k + \mu_k h_i(\theta_k)\big) \nabla h_i(\theta_k). \tag{9}$$

Comparing (9) with the KKT stationarity condition suggests defining

$$\nu_i^{k+1} := \nu_i^k + \mu_k h_i(\theta_k), \qquad i \in \mathcal{E}. \tag{10}$$

This motivates the multiplier update

$$\nu^{k+1} = \nu^k + \mu_k h(\theta_k). \tag{11}$$

Algorithm

Given ν^k, $\mu_k > 0$, and a starting point θ_k^s, the k-th ALM iteration is:

1. Primal step

$$\theta_k \approx \arg\min_{\theta} L_A(\theta, \nu^k; \mu_k), \qquad \|\nabla_\theta L_A(\theta_k, \nu^k; \mu_k)\| \le \tau_k.$$

2. Multiplier update

$$\nu^{k+1} = \nu^k + \mu_k h(\theta_k).$$

3. Parameter update

$$\theta_{k+1}^s \leftarrow \theta_k, \qquad \mu_{k+1} \ge \mu_k, \qquad \tau_{k+1} \text{ chosen as needed.}$$

Stopping criterion

$$\|h(\theta_k)\| \le \varepsilon_{\text{feas}}.$$

So each outer iteration has the pattern

$$\text{approximately minimize} \longrightarrow \text{update multiplier} \longrightarrow \text{adjust parameters.}$$

Summary of augmented Lagrangian method

- Unlike the quadratic penalty method, ALM improves the iterates not only by increasing μ, but also by updating the multiplier estimate ν^k.
- Therefore, good feasibility and accuracy can often be obtained with a moderate penalty parameter μ.
- A smaller μ helps avoid the severe ill-conditioning that appears in the quadratic penalty Hessian, so the subproblems are usually easier to solve.
- In this sense, ALM keeps the penalty idea, while reducing one of its main numerical drawbacks.
- Conceptually, ALM can be viewed as a stabilized dual-ascent method: the multiplier is still updated by the current constraint violation, while the primal step uses the augmented Lagrangian.