Lecture 20: CS395T Numerical Optimization for Graphics and AI — Penalty, Augmented Lagrangian and SDP

Qixing Huang
The University of Texas at Austin
huangqx@cs.utexas.edu

Disclaimer

This note is adapted from

-
- Section 17 of *Numerical Optimization* by Jorge Nocedal and Stephen J. Wright. Springer series in operations research and financial engineering. Springer, New York, NY, 2. ed. edition, (2006)
- http://mpc.zib.de/index.php/MPC/article/viewFile/40/20

1 The Quadratic Penalty Method

Consider the optimization problem described below:

$$\min_{\boldsymbol{x}} \ f(\boldsymbol{x}) \qquad \text{subject to } c_i(\boldsymbol{x}) = 0, \ i \in \mathcal{E}. \tag{1}$$

The quadratic penalty function $Q(\boldsymbol{x}, \mu)$ for this formulation is

$$Q(\boldsymbol{x}; \mu) := f(\boldsymbol{x}) + \frac{\mu}{2} \sum_{i \in \mathcal{E}} c_i^2(\boldsymbol{x}), \tag{2}$$

where $\mu > 0$ is the penalty parameter.

Quadratic Penalty Method.

- Given $\mu_0 > 0$, a non-negative sequence $\{\tau_k\}$ with $\tau_k \to 0$, and a starting point $\boldsymbol{x}_0^s$;
- **for** $k = 0, 1, 2, \cdots$
- Find an approximate minimizer $\boldsymbol{x}_k$ of $Q(\cdot; \mu_k)$, starting at $\boldsymbol{x}_k^s$, and terminating when $\|\nabla_x Q(\boldsymbol{x}; \mu_k)\| \leq \tau_k$
- if final convergence test satisfied **stop** with approximate solution $\boldsymbol{x}_k$; **end (if)**
- Choose new penalty parameter $\mu_{k+1} > \mu_k$;
- Choose new starting point x_{k+1}^s;
- **end (for)**

Convergence of The Quadratic Penalty Method.

Theorem 1.1. *Suppose that each $\boldsymbol{x}_k$ is the exact global minimizer of $Q(\boldsymbol{x};\mu_k)$ defined by (2) in the framework described above, and that $\mu_k \to \infty$. Then every limit point $\boldsymbol{x}^\star$ of the sequence $\boldsymbol{x}_k$ is a global solution of the problem.*

The theorem above considers the case where an exact solution is obtained at each iteration. For inexact solutions, we have:

Theorem 1.2. *Suppose that the tolerances and penalty parameters in the framework above satisfy $\tau_k \to 0$ and $\mu_k \to \infty$. Then if a limit point $\boldsymbol{x}^\star$ of the sequence $\{\boldsymbol{x}_k\}$ is infeasible, it is a stationary point of the function $\|c(\boldsymbol{x})\|^2$. On the other hand, if a limit point $\boldsymbol{x}^\star$ is feasible and the constraint gradients $\nabla c_i(\boldsymbol{x}^\star)$ are linearly independent, then $\boldsymbol{x}^\star$ is a KKT point for the problem. For such points, we have for any infinite subsequence $\mathcal{K}$ such that $\lim_{k\in\mathcal{K}} \boldsymbol{x}_k = \boldsymbol{x}^\star$ that*

$$\lim_{k\in\mathcal{K}} -\mu_k c_i(\boldsymbol{x}_k) = \lambda_i^\star, \qquad \text{for all } i \in \mathcal{E}.$$

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

2 Augmented Lagrangian Method: Equality Constraints

We consider first the equality-constrained problem. The quadratic penalty function $Q(\boldsymbol{x};\mu)$ penalizes constraint violations by squaring the infeasibilities and scaling them by $\mu/2$. As we see from Theorem 1.2, however, the approximate minimizers $\boldsymbol{x}_k$ of $Q(\boldsymbol{x};\mu_k)$ do not quite satisfy the feasibility conditions $c_i(\boldsymbol{x}) = 0, i \in \mathcal{E}$. Instead, they are perturbed so that

$$c_i(\boldsymbol{x}_k) \approx -\lambda_i^\star/\mu_k, \quad \text{for all } i \in \mathcal{E}. \tag{3}$$

To be sure, we have $c_i(\boldsymbol{x}_k) \to 0$ as $\mu_k \to \infty$, but one may ask whether we can alter the function $Q(\boldsymbol{x};\mu_k)$ to avoid this systematic perturbation-that is, to make the approximate minimizers more nearly satisfy the equality constraints $c_i(\boldsymbol{x}) = 0$, even for moderate values of μ_k. The augmented Lagrangian function $\mathcal{L}_A = (\boldsymbol{x},\lambda;\mu)$ achieves this goal by including an explicit estimate of the Lagrange multipliers λ, based on the estimate (3), in the objective. From the definition

$$\mathcal{L}_A(\boldsymbol{x},\lambda;\mu) := f(\boldsymbol{x}) - \sum_{i\in\mathcal{E}} \lambda_i c_i(\boldsymbol{x}) + \frac{\mu}{2}\sum_{i\in\mathcal{E}} c_i^2(\boldsymbol{x}). \tag{4}$$

we see that the augmented Lagrangian differs from the (standard) Lagrangian by the presence of the squared terms, while it differs from the quadratic penalty function (2) in the presence of the summation term involving λ. In this sense, it is a combination of the Lagrangian function and the quadratic penalty function. We now design an algorithm that fixes the penalty parameter μ to some value $\mu_k > 0$ at its kth iteration, fixes λ at the current estimate λ_k, and performs minimization with respect to $\boldsymbol{x}$. Using $\boldsymbol{x}_k$ to denote the approximate minimizer of $\mathcal{L}_A(\boldsymbol{x},\lambda_k;\mu_k)$, we have by the optimality conditions for unconstrained minimization that

$$0 \approx \nabla_{\boldsymbol{x}}\mathcal{L}_A(\boldsymbol{x}_k,\lambda_k;\mu_k) = \nabla f(\boldsymbol{x}_k) - \sum_{i\in\mathcal{E}}[\lambda_i^k - \mu_k c_i(\boldsymbol{x}_k)]\nabla c_i(\boldsymbol{x}_k). \tag{5}$$

By comparing with the optimality conditions (assuming x_k is already close to $\boldsymbol{x}^\star$, we can deduce that

$$\lambda_i^\star \approx \lambda_i^k - \mu_k c_i(\boldsymbol{x}_k), \quad \text{for all } i \in \mathcal{E}. \tag{6}$$

By rearranging this expression, we have that

$$c_i(\boldsymbol{x}_k) \approx -\frac{1}{\mu_k}(\lambda^\star - \lambda^k), \quad \text{for all } i \in \mathcal{E},$$

so we conclude that if λ_k is close to the optimal multiplier vector $\lambda^\star$, the infeasibility in $\boldsymbol{x}^k$ will be much smaller than $1/\mu_k$, rather than being proportional to $1/\mu_k$ as in Theorem 1.2. The relation (6) immediately suggests a formula for improving our current estimate λ_k of the Lagrange multiplier vector, using the approximate minimizer $\boldsymbol{x}_k$ just calculated: We can set

$$\lambda_i^{k+1} = \lambda_i^k - \mu_k c_i(\boldsymbol{x}_k), \quad \text{for all } i \in \mathcal{E}.$$

Augmented Lagrangian Method-Equality Constraints.

- Given $\mu_0 > 0$, tolerance $\tau_0 > 0$, and a starting point $\boldsymbol{x}_0^s$ and λ^0;
- **for** $k = 0, 1, 2, \cdots$
- Find an approximate minimizer $\boldsymbol{x}_k$ of $\mathcal{L}_A(\cdot, \lambda^k; \mu_k)$, starting at $\boldsymbol{x}_k^s$, and terminating when
$$\|\nabla_x \mathcal{L}_A(\boldsymbol{x}_k, \lambda^k; \mu_k)\| \le \tau_k$$
- if final convergence test satisfied **stop** with approximate solution $\boldsymbol{x}_k$; **end (if)**
- Update Lagrange multipliers using (6) to obtain λ^{k+1};
- Choose new penalty parameter $\mu_{k+1} > \mu_k$;
- Choose new starting point $x_{k+1}^s = x_k$;
- Select tolerance τ_{k+1};
- **end (for)**

Properties of The Augmented Lagrangian. We now prove two results that justify the use of the augmented Lagrangian function and the method of multipliers for equality-constrained problems. The first result validates the approach of above framework by showing that when we have knowledge of the exact Lagrange multiplier vector $\lambda^\star$ the solution $\boldsymbol{x}^\star$ is a strict minimizer of $\mathcal{L}_A(\boldsymbol{x}, \lambda; \mu)$ for all μ sufficiently large. Although we do not know $\lambda^\star$ exactly in practice, the result and its proof suggest that we can obtain a good estimate of $\boldsymbol{x}^\star$ by minimizing $\mathcal{L}_A(\boldsymbol{x}, \lambda; \mu)$ even when μ is not particularly large, provided that λ is a reasonably good estimate of $\lambda^\star$.

Theorem 2.1. *Let $\boldsymbol{x}^\star$ be a local solution of (1) at which the LICQ is satisfied (that is, the gradients$\nabla c_i(\boldsymbol{x}^\star), i \in \mathcal{E}$, are linearly independent vectors), and the second-order sufficient conditions specified are satisfied for $\lambda = \lambda^\star$. Then there is a threshold value μ^- such that for all $\mu \ge \mu^-$, $\boldsymbol{x}^\star$ is a strict local minimizer of $\mathcal{L}_A(\boldsymbol{x}, \lambda^\star; \mu)$.*

Theorem 2.2. *Suppose that the assumptions of Theorem 2.1 are satisfied at $\boldsymbol{x}^\star$ and $\lambda^\star$ and let μ^- be chosen as in that theorem. Then there exist positive scalars δ, ϵ and M such that the following claims hold:*

- *For all λ_k and μ_k satisfying*
$$\|\lambda^k - \lambda^\star\| \le \mu_k \delta, \quad \mu_k \ge \mu^-, \tag{7}$$
the problem
$$\min_{\boldsymbol{x}} \mathcal{L}_A(\boldsymbol{x}, \lambda_k; \mu_k) \quad s.t. \|\boldsymbol{x}^\star - \boldsymbol{x}^\star\| \le \epsilon$$
has a unique solution $\boldsymbol{x}_k$. Moreover, we have
$$\|\boldsymbol{x}_k - \boldsymbol{x}^\star\| \le M \|\lambda_k - \lambda^\star\| / \mu_k.$$
- *For all λ_k and μ_k that satisfy (7), we have*
$$\|\lambda_{k+1} - \lambda^\star\| \le M \|\lambda_k - \lambda^\star\| / \mu_k, \tag{8}$$
where λ_{k+1} is given by the formula (6).
- *For all λ_k and μ_k that satisfy (7), the matrix $\nabla^2_{\boldsymbol{xx}} \mathcal{L}_A(\boldsymbol{x}_k, \lambda_k; \mu_k)$ is positive definite and the constraint gradients $\nabla c_i(\boldsymbol{x}_k), i \in \mathcal{E}$, are linearly independent.*