Lecture 3: Continuous spaces

1 General problem

$$\min_{x \in \mathcal{R}^n} f(x)$$

subject to $c_i(x) = 0$ and $d_i(x) > 0$.

1.1 Example problems

- Fitting a model to fixed data set
- Placing fire stations
- Allocating investments

1.2 Assumptions we can take advantage of

- f is linear
- f is quadratic
- f is smooth
- c and d are linear

When modeling: trade off simplicity of model (for ease of solution) against its goodness of fit to the problem.

2 Unconstrained problems

If f is linear, then there's no solution.

If it's quadratic, and convex, then it's easy.

More generally, we have to do local optimization.

$$x_{k+1} = x_k + \alpha_k p_k$$

where α_k is a step size and p_k is a direction.

2.1 Pick a direction

Gradient descent (or steepest descent):

$$p_k = -\nabla f(x_k)$$

Newton direction:

$$p_k = -(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$$

Exact if f is quadratic. Typically very expensive to deal with Hessian ($\delta^2 f$).

Conjugate gradient: not as good as Newton, but much better than gradient and much more efficient than Newton. Very popular. Tries to find a new direction that is conjugate to previous one. Linear number of steps if f is quadratic.

$$p_{k+1} = -\nabla f(x_k) + \beta_{k+1} \cdot p_k$$

where

$$\beta_{k+1} = \frac{\nabla f(x_{k+1})^T \nabla f(x_{k+1})}{\nabla f(x_k)^T \nabla f(x_k)}$$

Calculating derivatives If you don't know the derivative analytically, then you can do a simple finite difference. Pick ϵ to be small, but be careful about roundoff.

$$\frac{\partial f}{\partial x_i}(x) \approx \frac{f(x + \epsilon e_i) - f(x)}{\epsilon}$$

where e_i is the ith unit vector.

2.2 Pick a step size

Simple gradient descent uses a fixed α. Tricky to find a good one (too big can oscillate, too small can be slow).

More effectively, once you have picked a direction, is to do a *line search* in that direction to find a good step size. Could try to find the best, but usually stop after finding one that's 'good enough' based on theoretical criteria.

3 Constrained problems

3.1 Linear programs

Both f and the constraints are linear functions of x. Possible situations:

- infeasible (constraint region is empty)
- unbounded (insufficiently constrained in the direction that f is growing)
- due to perfect alignment, there is a set of solutions along a face or edge of the constraint simplex
- single solution at a vertex

Simplex method: considers vertices. Usually efficient but worst case exponential in the number of dimensions.

Interior point methods: worst case polynomial in dimenions (but early versions were usually much worse than simplex on actual problems).

3.2 Quadratic programs

Linear constraints, but quadratic f. Relatively easy if f is convex. Also easier to deal with equality constraints.

3.3 General form of function and constraints

Convert to an unconstrained optimization problem by added in a penalty:

$$Q(x;\mu) = f(x) + \mu \sum_i c_i^2(x) + \mu \sum_i ([d_i(x)]^-)^2$$

Gradually increase μ. Changes minimizer.

Method of multipliers allows us to find a single optimization problem, with extra parameters, whose minimizer is a minimizer of the original problem.

3.4 Sequential quadratic programming

Fit a local quadratic model; solve via QP; take a step; repeat.

4 Uncertainty

4.1 Stochastic gradient

In machine learning, we don't always know the f we're trying to optimize. For instance, we might want a function to fit well in expectation, but we can only draw samples.

$$f(x) = \sum_s Pr(s) g(s;x)$$

Gradient is also an expectation:

$$\frac{df}{dx}(x) = \sum_s Pr(s) \frac{dg}{dx}(s;x)$$

Two strategies:

- Draw a lot of samples of s to get a good estimate of the gradient, and then take a step.
- Draw a single sample and take a step.

The second approach can work, but need to use a shrinking step size (usually as $1/k$).

4.2 Response surface methodology

We don't know f, so we draw samples, fit a surface $\hat{f}$, and take a gradient step with respect to $\hat{f}$, get a new point, fit a new surface, etc.