10-725: Optimization **Fall 2012**

Lecture 23: November 15

Lecturer: Geoff Gordon/Ryan Tibshirani *Scribes: Jing Xiang*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

23.1 Review: SVMs

Last lecture was about support vector machines. Shown below is the primal dual pair for SVMs. The primal is to maximize the margin, whereas the dual is a minimization problem. K is the gram matrix, the matrix of pairwise dot products. The gram matrix helps you solve kernelized SVMs really quickly even in high dimensions.

$$
\begin{aligned}
&\min \frac{v^\top v}{2} + c1^\top s \text{ s.t } Av - yd + s - 1 \geq 0, s \geq 0 \\
&\max 1^\top \alpha - \frac{\alpha^\top K \alpha}{2} \text{ s.t. } y^\top \alpha = 0, 0 \leq \alpha \leq 1 \\
&K \text{ gram matrix}
\end{aligned}
$$

We can use the complementarity conditions to reconstruct the primal from the dual. To get the intercept, we use complementary slackness.

The kernel trick is how to do high dimensional feature spaces fast. It works for any positive definite function. Some examples are polynomials, homogeneous polynomials, linear functions, and Gaussian RBFs.

To avoid overfitting, kernels use large margins. There are not many functions that can separate your data with large margins. Also, big penalties on slacks means not very complicated functions. If you make c large enough, it will pick v equal to 0, you will not overfit if you don't fit. If you make c small, then you get a flexible kernel and it will overfit.

23.2 Review: Linear Feasibility Problem

- Ball center
  - Ball center can be a bad summary, ex. skinny long triangle.
- Max-volume ellipsoid/ellipsoid center
  - This is a much better summary ($\frac{1}{N^N}$ of volume) but it is much more expensive to find it.
- Analytic center of LF problem

  - The point that maximizes the product of distances to the constraints. You can think of this as a soft version of the ball center. If one of the distances is very small, it will make the whole product small.
  - $\min -\sum \log(a_i^\top x + b_i)$

- Dikin ellipsoid at the analytic center
  - This is a pretty good summary, can prove that it contains $\frac{1}{M^N}$ of the volume. This is not as good as $\frac{1}{N^N}$ but is cheaper to compute.

23.3 Analytic Center

One interpretation of the analytic center is to pretend that each constraint is repelling the particle. So, there is a normal force for each constraint. The strength of the force is $\propto 1/\text{distance}$. Where the forces balance is where the analytic center is.

To find the analytic center, you can use the Newton's method. Define $s_i = a_i^\top x + b_i$, $y_i = \frac{1}{s_i}$ and $S = diag(s)$.

$$\begin{aligned} f(x) &= -\sum_i \log(a_i^\top x + b_i) \\ \frac{df}{dx} &= -\sum_i \left(\frac{1}{a_i^\top x + b_i}\right) a_i \\ 0 &= -A^\top y \text{ First Order Condition} \end{aligned}$$

$$\begin{aligned} \frac{d^2 f}{dx^2} &= \sum_i (a_i^\top x + b_i)^{-2} a_i a_i^\top \\ H &= A^\top S^{-2} A \\ H &= A^\top Y^2 A \end{aligned}$$

The Newton step is then $\Delta x = (A^\top S^{-2} A)^{-1} A^\top y$.
This is a strictly convex function which means that the Hessian is going to be strictly positive definite. The Newton direction is going to be a descent direction. So Newton with line search will converge from any strictly feasible initial point. In fact, it will converge quite quickly.

23.4 Dikin ellipsoid

The Dikin ellipsoid centered at x_0 can be written as $E(x_0) = \{x | (x - x_0)^\top H (x - x_0) \le 1\}$. H here is the Hessian of the log barrier function at x_0. Another way of saying this is that the Dikin ellipsoid is the unit ball of the Hessian norm at x_0. A diagram of the Dikin ellipsoid is shown below (Fig. 23.1).

We can prove two things about the Dikin ellipsoid. For any strictly feasible x_0, if we calculate the Dikin ellipsoid centered at that point, that's going to be contained in the feasible region ($E(x_0) \subseteq X$). This means

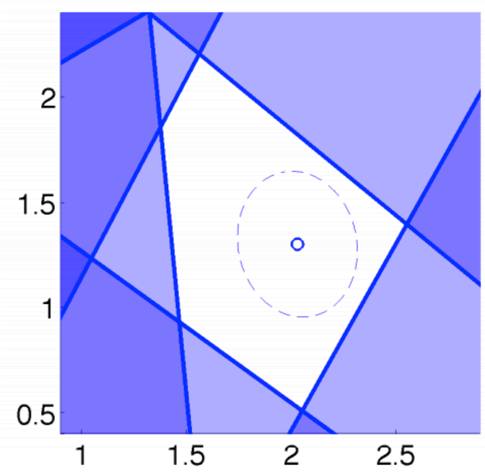

Figure 23.1: Diagram of the Dikin ellipsoid

that the Dikon ellipsoid doesn't just let us summarize the geometry of the feasible region, it allows us to summarize the geometry of the feasible region near a particular point.

Show that $E(x_0) \subseteq X$, the Dikin ellipsoid centered at x_0 is contained in the feasible region.

$E(x_0) = \{x|(x - x_0)^\top H(x - x_0) \leq 1\}$
$H = A^\top S^{-2} A$
$S = diag(s) = diag(Ax_0 + b)$

$$
\begin{aligned}
&\frac{\sum_i (a_i^\top (x + x_0))^2}{s_i} \leq 1 \\
&-1 \leq \frac{a_i^\top (x + x_0)}{s_i} \leq 1 \ \forall i \\
&-s_i \leq a_i^\top (x - x_0) \\
&0 \leq a_i^\top x + b_i
\end{aligned}
$$

You can also take the ellipsoid centered at the analytic center, and if you grow it by m at the center, then it contains the feasible region. This means it doesn't just tell us an inscribed ellipsoid, it tells us a circumscribed ellipsoid so we have inner and outer bounds on the feasible region from the same calculation. The fact that you have to scale it by a factor of m, means that the original ellipsoid must have at least 1/m of the volume of the feasible region.

Show that $mE(x_{ac}) \subseteq X$, the Dikin ellipsoid expanded by a factor of m centered at the analytic center is contained in the feasible region. First define the following:

Feasible point x: $Ax + b \geq 0$
Analytic center x_{ac}: $A^\top y = 0$, $y = 1./(Ax_{ac} + b)$
Let $Y = diag(y_{ac})$, $H = A^\top Y^2 A$ and show that:

$(x - x_{ac})^\top H(x - x_{ac}) \leq m^2$ $[+m]$

$$\begin{aligned}
\sum_i y_i^2(a_i^\top(x - x_{ac}))^2 &= y_i^2(a_i^\top x + b_i - \frac{1}{y_i}))^2 \\
&= \sum_i y_i^2((a_i^\top x + b_i)^2 - 2(a_i^\top x + b_i)y_i + \frac{1}{y_i})) \\
&\leq m + \sum_i y_i^2(a_i^\top x + b_i)^2 \leq m + (\sum_i y_i(a_i^\top x + b_i))^2 \\
&\leq m^2 + m
\end{aligned}$$

There are two very different ways to find a feasible point of your set of linear inequalities. One is to find a feasible basis which is a combinatorial search and the other is to find an analytic center by minimizing a smooth function. So, linear programs sits halfway between the analysis world and the combinatoric world.

23.5 Bad Conditioning

One of the nice things about the analytic center and the Dikin ellipsoid is that it relieves bad conditioning. If you have a skinny triangular feasible region (Fig. 23.2), the Dikin ellipsoid still takes up a pretty big volume of the feasible region.

You can imagine taking an affine transformation of the feasible region, to take the analytic center to the origin. The analytic center and Dikin ellipsoid is invariant to affine transforms.

$w = Mx + q$
$W = \{w | AM^{-1}(wq) + b \geq 0\}$

We assume that H_x factors into the Cholesky decomposition $R^\top R$ for a lower triangular R. If we take $M = R$.

$$\begin{aligned}
w_{ac} &= Mx_{ac} + q \\
H_w &= M^{-T} H_x M^{-1} \\
H_w &= R^{-T} R^T R R^{-1} \\
&= I
\end{aligned}$$

So we can always take the Dikin ellipsoid to a sphere and the analytic center to the origin by an appropriate affine transformation. We expect this because we know that an affine transformation is not going to change where the optimum of a convex function is and we know that Newton is affine invariant.

23.6 Central Path from LF to LP

The analytic center was to find x s.t. Ax + b $\geq$ 0.

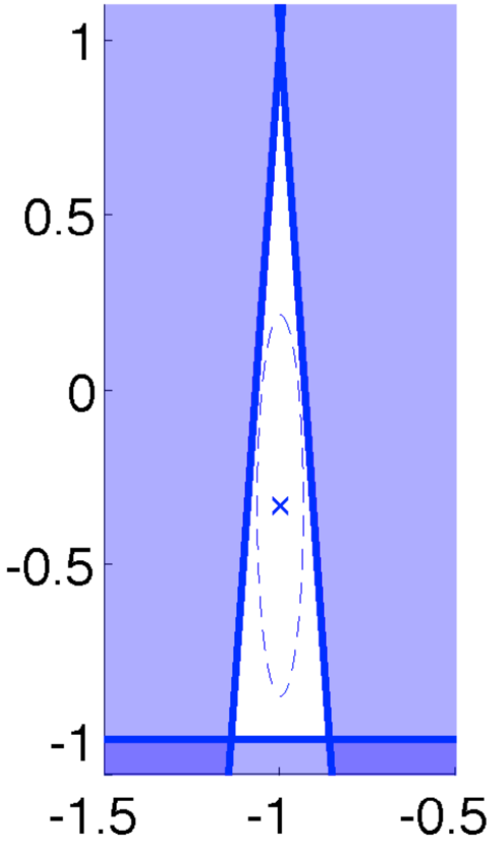

Figure 23.2: Diagram Ill Conditioned LP.

For a linear program, instead of finding x, we want to do the following.
Now we want to $\min c^\top x$ s.t. $Ax + b \geq 0$

We can use the same trick. Here we have a centering force and an optimization force. We have a parameter t that trades between the strength of the two forces. If t approaches zero, then 1/0 approaches infinity, the barrier term becomes important so you get the analytic center and as t approaches infinity, then you get the LP optimum. The central path has at one end, the analytic center and at the other, the LP optimum. In fact, the central path is smooth and the gradient of the central path is related to the Newton direction ($\propto H^{-1}c$). The central path is also invariant to affine transformations.

- $\min f_t(x) = c^\top x \frac{1}{t} \sum \log(a_i^\top x + b_i)$
- parameter $t > 0$
- central path $= \{x(t) | t > 0\}$
- $t \to 0$: analytic center, $t \to \infty$: LP optimum

There is a force field interpretation for the central path (Fig. 23.4). We have a force for trying to go in the direction of the objective function. The higher t is, the longer the arrow is in the direction of that force. So,

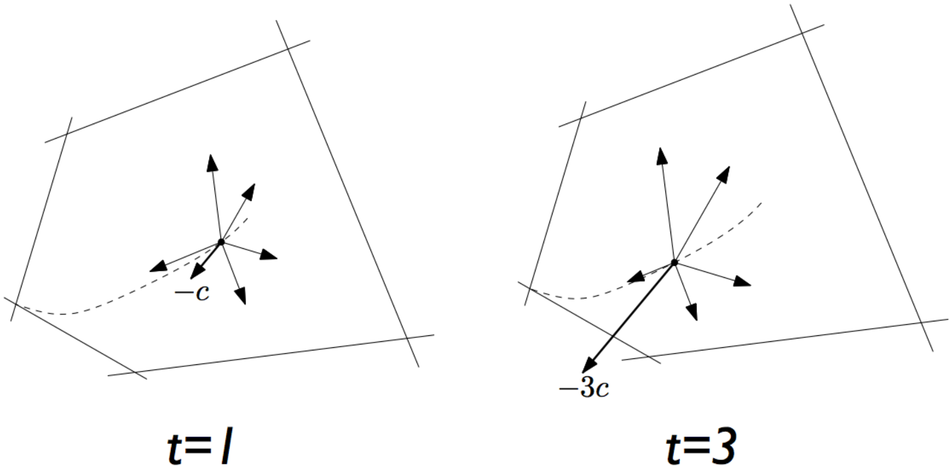

Figure 23.3: Forces along objective, normal forces for each constraint.

there will be a different trade off between repulsive forces of the barrier and the optimality force from the objective. The central path point is the equilibrium point where the forces balance.

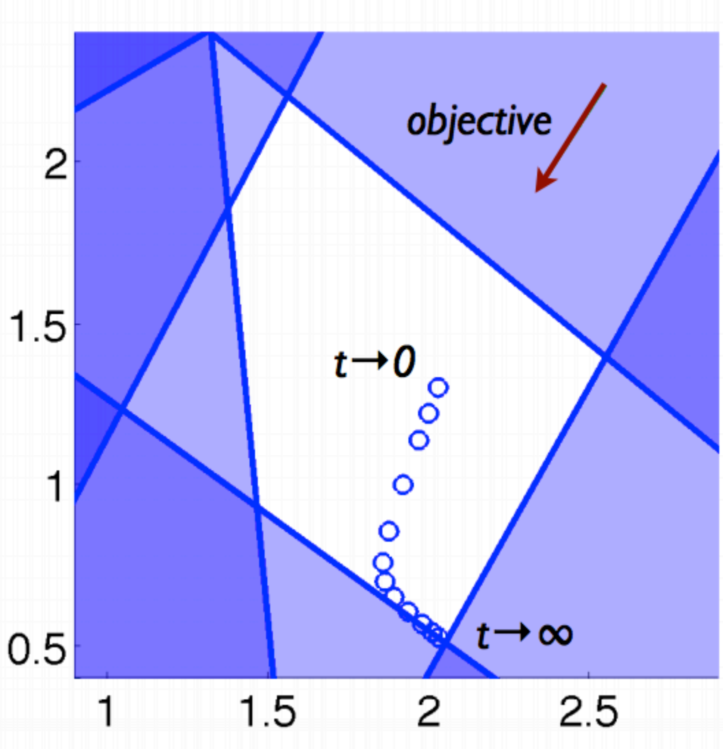

Figure 23.4: Central path example.

23.7 Newton's Method for the central path

$$\begin{aligned}
&\min f_t(x) = c^\top x - \frac{1}{t}\sum \log(a_i^\top x + b_i) \\
&\frac{df}{dx} = c - \frac{1}{t}\sum \frac{1}{s_i} a_i \\
&\frac{d^2 f}{dx^2} = \frac{1}{t}\sum s_i^{-2} a_i a_i^\top
\end{aligned}$$

Note that $\sum s_i^{-2} a_i a_i^\top \Delta x = \sum \frac{a_i}{s_i} - tc$ is the negative of the gradient and the $-tc$ term is different from the Newton direction for the analytic center. If t is small, we get the analytic center.

23.8 Dual of central path

We can also take the dual of the central path problem.

$$\begin{aligned}
&\min c^\top x - (1/t)\sum \log s_i \text{ s.t. } Ax + b = s \geq 0 \\
&\min_{x,s} \max_{y} L(x, s, y) = c^\top x - \frac{1}{t}\sum \log s_i + y^\top (s - Ax - b) \\
&\Delta_x = c - A^\top y \text{ so } c = A^\top y \\
&\Delta_s = -\frac{1}{ts} + y \text{ so } s = \frac{1}{ty}
\end{aligned}$$

$$\begin{aligned}
&\max_y \frac{m}{t} - y^\top b + \frac{1}{t}\sum (\log y_i + \log t \\
&\max_y \frac{m}{t} + \frac{m \log t}{t} - y^\top b + \frac{1}{t}\sum \log y_i \\
&\text{s.t. } A^\top y = c \ (y \geq 0)
\end{aligned}$$

Why is this interesting? One thing is that it is easy to find a point that is strictly feasible for the inequality constraints. You can have $y_i = 1$ for all y. You can then an use an infeasible start Newton's method. Start with a y that doesn't satisfy equality constraints. But we know how to make Newton's method satisfy equality constraints over time. So we can run this from a trivial initializer. That means you don't have to worry about finding a feasible point to start. This gives you a slightly different Newton update. It is a slightly different algorithm which has very simpler behavior and guarantees.

23.9 Primal Dual correspondance

The other thing that is cool about the dual is that there is a correspondence between the primal/dual central paths. There is a strong correspondence between every point in the primal central problem and the corresponding point in the dual.

$$L(x,s,y) = c^\top x \frac{1}{t} \sum \log s_i + y^\top (s - Ax - b)$$

$$\text{The gradient w.r.t. s: } y - \frac{\mathbf{1}}{t} \Rightarrow s_i y_i = \frac{\mathbf{1}}{t} \Rightarrow s \circ y = \frac{\mathbf{1}}{t}$$

$$\text{To get x: } Ax + b = \frac{\mathbf{1}}{ty}$$

23.10 Duality Gap

At the optimum the primal value will equal the dual value. The duality gap is just $\frac{m}{t}$ if we pick t as the parameter.

Note that $\sum \log s_i y_i = -\sum \log t$ and
$\sum(\log s_i + \log y_i) = -m \log t$

$$c^\top x - \frac{1}{t} \sum \log s_i = \frac{m \log t}{t} + \frac{m}{t} + \frac{1}{t} \sum \log y_i - y^\top b \qquad c^\top x + y^\top b = \frac{m}{t}$$