STAT 538 Final Exam Solutions
Friday March 5, 2010, 3:30-5:20

Problem 1 – Maxima of convex functions

1.1 Assume that x^* is not an extreme point. Then there are $x_1, x_2 \in C$, $x_1, x_2 \neq x^*$, so that $x^* = tx_1 + (1-t)x_2$ for some $t \in (0,1)$. Then,

$$f(x^*) \;<\; tf(x_1) + (1-t)f(x_2) \;\leq\; tf(x^*) + (1-t)f(x^*) \;=\; f(x^*) \tag{1}$$

We have arrived at a contradiction, hence x^* must be an extreme point.

1.2 x^* is not unique. Counterexample: $f(x) = x^2 - 1$, $C = [-1, 1]$; f has two maxima at 1 and -1.

1.3 x^* is not isolated. Counterexample in $\mathbb{R}^n$: $f(x) = ||x||^2 - 1$, $C = \{||x|| \leq 1\}$. Every point of the boundary of the unit ball is a maximum of f, and an extreme point of C.

Problem 2 – The rate of convergence of gradient descent with line minimization

2.1

$$\begin{aligned} g &= \nabla f \;=\; Qx & (2)\\ f(x - \alpha g) &= \frac{1}{2}x^T Q x + \frac{\alpha^2}{2} g^T Q g - \alpha x^T Q g & (3)\\ \frac{d}{d\alpha} f(x - \alpha g) &= \alpha g^T Q g - x^T Q g \;=\; 0 & (4)\\ \alpha &= \frac{x^T Q g}{g^T Q g} \;=\; \frac{g^T g}{g^T Q g} & (5) \end{aligned}$$

The latter equality follows because $x = Q^{-1} g$.

2.2

$$f(x - \alpha g) \;=\; \frac{1}{2}\left[x^T Q x - \frac{(x^T Q g)^2}{g^T Q g}\right] \tag{6}$$

$$
\begin{aligned}
\frac{f(x-\alpha g)}{f(x)} &= \frac{x^TQx - \frac{(x^TQg)^2}{g^TQg}}{x^TQx} &\quad (7)\\
&= 1 - \frac{(x^TQg)^2}{g^TQgx^TQx} &\quad (8)\\
&= 1 - \frac{(g^Tg)^2}{(g^TQg)(g^TQ^{-1}g)} &\quad (9)
\end{aligned}
$$

The latter equality follows because $x = Q^{-1}g$.

2.3 First, we get the eigenvalues of Q:

$$
\begin{vmatrix} \lambda - 2 & -a \\ -a & \lambda - 2 \end{vmatrix} = (\lambda-2)^2 - a^2 = (\lambda-\epsilon)(\lambda - 4 + \epsilon)
$$

It follows that $\lambda_1 = m = \epsilon, \lambda_2 = M = 4 - \epsilon$. Hence,

$$
\frac{f(x-\alpha g)}{f(x)} \le 1 - \frac{4mM}{(M+m)^2} = 1 - \frac{4\epsilon(4-\epsilon)}{4^2} = 1 - \epsilon(1-\epsilon/4)
$$

For small ϵ, this rate is nearly 1, and convergence will be very slow.

Problem 3 – SVM with logarithmic penalty

$$
\begin{aligned}
(\mathcal{P}) \quad \min_{w,b,\gamma_{1:m}} \quad & \tfrac{1}{2}||w||^2 + \textstyle\sum_{i=1}^m \ln \frac{1+e_i^\gamma}{2} \\
\text{s.t.} \quad & y_i(w^Tx_i + b) \ge 1 - \gamma_i, \text{ for all } i &\quad (10)\\
& \gamma_i \ge 0 \text{ for all } i &\quad (11)
\end{aligned}
$$

3.1 $\gamma_{1:m}$ are called *slack variables*. Their role is to measure the amount by which the margin conditions (2) are violated in the solution. $\gamma_i = 0$ whenever a point is classified with margin 1 or larger.

3.2, 3.3

$$
\begin{aligned}
L(w,b,\gamma,\lambda,\alpha) &= \frac{1}{2}||w||^2 + \sum_{i=1}^m \ln\frac{1+e_i^\gamma}{2} + \sum_i \lambda_i \left[1 - \gamma_i - y_i(w^Tx_i + b)\right] - \sum_i \alpha_i\gamma_i \\
\frac{\partial L}{\partial w} &= w - \sum_i \lambda_i y_i x_i \quad \Rightarrow \quad w = \sum_i \lambda_i y_i x_i &\quad (12)\\
\frac{\partial L}{\partial b} &= \sum_i \lambda_i y_i &\quad (13)\\
\frac{\partial L}{\partial \gamma_i} &= \frac{e_i^\gamma}{1+e_i^\gamma} - \lambda_i - \alpha_i &\quad (14)\\
\gamma_i &= -\ln\left(\frac{1}{\lambda_i + \alpha_i} - 1\right) &\quad (15)
\end{aligned}
$$

It follows that $0 < \alpha_i + \gamma_i < 1$. Denote $K = [K_{ij}]$, $K_{ij} = y_i y_j x_i^T x_j$, $\beta_i = \alpha_i + \lambda_i$. Then

$$
\begin{aligned}
1 + e_i^\gamma &= \frac{1}{1-\beta_i} && (16)\\
g(\lambda, \alpha) &= \frac{1}{2}\lambda^T K \lambda - \sum_i \ln(1-\beta_i) - m \ln 2 - \sum_i \lambda_i + \sum_i \lambda_i \ln\left(\frac{1}{\beta_i} - 1\right) \\
&\quad - \left(\sum_i \lambda_i y_i x_i\right)^T \left(\sum_i \lambda_i y_i x_i\right) - \sum_i \alpha_i \ln\left(\frac{1}{\beta_i} - 1\right) && (17)\\
&= -\frac{1}{2}\lambda^T K \lambda - m \ln 2 - \sum_i \lambda_i + \sum_i \left[-\ln(1-\beta_i) + \beta_i \ln \frac{1-\beta_i}{\beta_i}\right] && (18)\\
&= -\frac{1}{2}\lambda^T K \lambda - m \ln 2 - \sum_i \lambda_i + \sum_i H(\beta_i) && (19)
\end{aligned}
$$

In the above $H(\beta_i)$ denotes the entropy $-\beta_i \ln \beta_i - (1-\beta_i)\ln(1-\beta_i)$.

3.4 It is easy to verify that g is concave: the term $-\lambda^T K \lambda$ is a negative quadratic, with K positive definite, the second and third terms are constant, respectively linear, and the terms $H(\beta_i)$ are entropies and therefore concave. The domain of the dual objective is $\lambda_i \in \mathbb{R}, \beta \in (0,1)$, convex.

$$
\begin{aligned}
(\mathcal{D}) \max_{\lambda,\alpha} \quad & -\tfrac{1}{2}\lambda^T K \lambda - m \ln 2 - \textstyle\sum_i \lambda_i + \sum_i H(\beta_i) \\
\text{s.t} \quad & \lambda_i \ge 0 && (20)\\
q \quad & \beta_i \ge \lambda_i && (21)\\
& \lambda^T y = 0 && (22)
\end{aligned}
$$

All constraints are linear, hence $(\mathcal{D})$ is a concave maximimization problem.

3.5 $(\mathcal{D})$ is not a quadratic problem, because of the entropy term $H(\beta_i)$.

3.6 $w^* = \sum_i \lambda_i^* y_i x_i$. Find an i for which $\lambda_i > 0$ Hence, $y_i(w^* T x_i + b) = 1 - \gamma_i^* = 1 - \ln \frac{1-\beta_i^*}{\beta_i^*}$. From this equation, $b^* = y^i(1-\gamma_i^*) - w^{*T} x_i$. The resulting classifier is $f(x) = (x^T w^* + b^*)$.

3.7 If $y_i(x_i^T w + b) < 1$ then $\gamma_i > 0$, then the corresponding $\alpha_i = 0$ by complementary slackness, and $\lambda_i > 0$ because the constraint (10) is tight.

3.8 If $y_i(x_i^T w + b) > 1$ then $\gamma_i = 0$ and the constraint (10) is slack, while the constraint (11) is tight. Hence the corresponding dual variables are $\alpha_i > 0$ and $\lambda_i = 0$, by complementary slackness. For $\gamma_i = 0$ it follows that $\beta_i = \frac{1}{2} = 0 + \alpha_i$, hence $\alpha_i = 1/2$.

3.9 Let (x_i, y_i) be a data point for which (w^*, b^*) has margin $= 1$. What can you say about λ_i, γ_i, α_i in this case? Find λ_i as a function of the other λ's.

For $y_i(x_i^T w + b) = 1$ we have $\gamma_i = 0$, $\beta_i = 1/2$ as above, and $\alpha_i > 0$ typically. Hence $\lambda_i + \alpha_i = 1/2$. The dual objective g can be written as

$$g = -\lambda_i^2 K_{ii}/2 - \underbrace{\frac{1}{2}\sum_{j \neq i} \lambda_j K_{ij}}_{k_i} \lambda_i + H(\beta_i) + \text{terms independent of } \lambda_i$$

Also, $H(\beta_i) = \log 2$ for any λ_i. So the maximum over λ_i is attained for

$$\lambda_i = \begin{cases} -k_i/K_{ii} & \text{if } k_i/K_{ii} \in (-1/2, 0) \\ 0, & \text{if } k_i \geq 0 \\ \frac{1}{2}, & \text{otherwise} \end{cases}$$