CPSC 340:
Machine Learning and Data Mining

Gradient Descent

BONUS SLIDES

Bonus Slide: Invertible Matrices and Regularization

- Unlike least squares where X^TX may not be invertible, the matrix $(X^TX + \lambda I)$ in always invertible.

- We prove this by showing that $(X^TX + \lambda I)$ is positive-definite, meaning that $v^T(X^TX + \lambda I)v > 0$ for all non-zero 'v'. (Positive-definite matrices are invertible.)

With a generic 'v' such that $v \neq 0$ we have

$$v^T(X^TX + \lambda I)v = v^TX^TXv + \lambda v^Tv$$

$$= \underbrace{\|Xv\|^2}_{\geq 0} + \underbrace{\lambda \sum_{j=1}^{d} v_j^2}_{> 0 \text{ since } v \neq 0.}$$

Bonus Slide: Log-Sum-Exp for Brittle Regression

- To use log-sum-exp for brittle regression:

$$\|Xw - y\|_\infty = \max_i \{|w^T x_i - y_i|\}$$

$$= \max_i \{\max\{w^T x_i - y_i, y_i - w^T x_i\}\}$$ since $|z| = \max\{z, -z\}$

$$= \log\left(\sum_{i=1}^n \exp(w^T x_i - y_i) + \sum_{i=1}^n \exp(y_i - w^T x_i)\right)$$ using log-sum-exp to approximate "max" over $2n$ terms.

Bonus Slide: Log-Sum-Exp Numerical Trick

- Numerical problem with log-sum-exp is that $\exp(z_i)$ might overflow.
  – For example, exp(100) has more than 40 digits.
- Implementation 'trick': Let $\beta = \max_i \{z_i\}$

$$
\begin{aligned}
\log\left(\sum_i \exp(z_i)\right) &= \log\left(\sum_i \exp(z_i - \beta + \beta)\right) \\
&= \log\left(\sum_i \exp(z_i - \beta)\exp(\beta)\right) \\
&= \log\left(\exp(\beta) \sum_i \exp(z_i - \beta)\right) \\
&= \log(\exp(\beta)) + \log\left(\sum_i \exp(z_i - \beta)\right) \\
&= \beta + \log\left(\sum_i \exp(z_i - \beta)\right)
\end{aligned}
$$

≤ 1 so no overflow

Bonus Slide: Normalized Steps

Question from class: "can we use $w^{t+1} = w^t - \frac{1}{\|\nabla f(w^t)\|}\nabla f(w^t)$"

This will work for a while, but notice that

$$\|w^{t+1} - w^t\| = \left\|\frac{1}{\|\nabla f(w^t)\|}\nabla f(w^t)\right\|$$
$$= \frac{1}{\|\nabla f(w^t)\|}\|\nabla f(w^t)\|$$
$$= 1$$

So the algorithm <u>never converges</u>

Bonus Slide: Gradient Descent for Non-Smooth?

- "You are unlikely to land on a non-smooth point, so gradient descent should work for non-smooth problems?"
  - Counter-example from Bertsekas' "Nonlinear Programming"

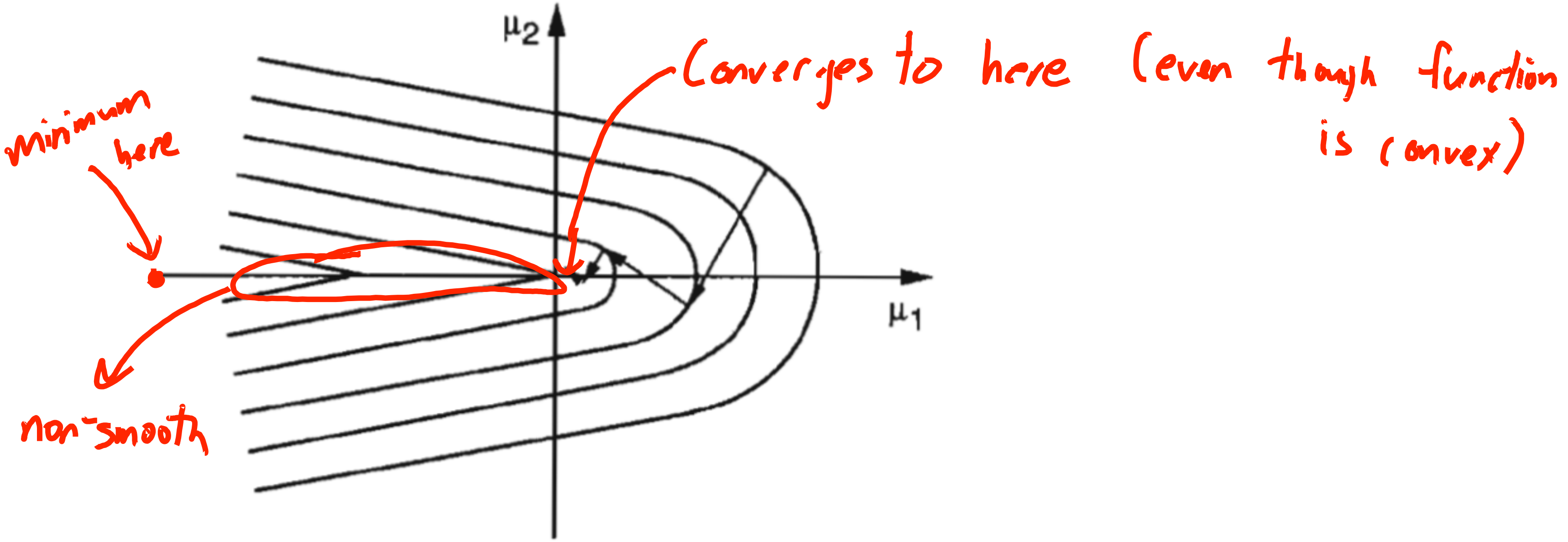

Figure 6.3.8. Contours and steepest ascent path for the function of Exercise 6.3.8.