Tutorial 6

Convexity and Regularization

Adapted from Issam Laradji's Slides

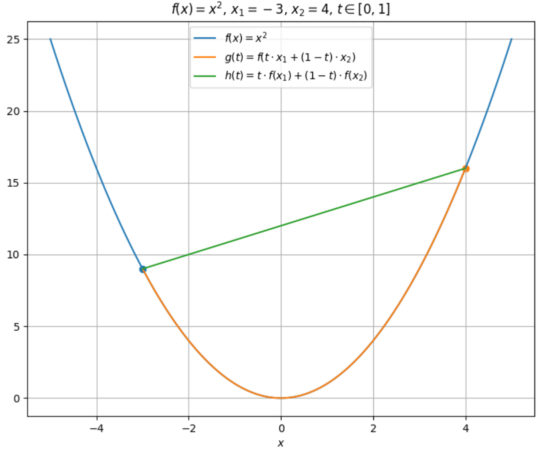
Outline

- Convex Functions
- Regularization
- Assignment Code

Definition of convexity: Jensen's Inequality

- A function f is convex if $\forall x_1, x_2 \in \mathbb{R}; \forall t \in [0, 1]$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$



Intuition and Proofs

- Why do we like convex functions?
  - Hint: What does Jensen's inequality say about optima?
- How do we prove that functions are convex?

1. Linear functions are convex

- $f(x) = Ax$ is a convex function
  - where A is some 2D matrix in $\mathbb{R}$
- **proof.**
  - A function f is convex if for $\forall x_1, x_2 \in \mathbb{R}; \forall t \in [0, 1]$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

  - By definition, a linear function is:

$$\begin{aligned} f(tx_1 + (1-t)x_2) &= A(tx_1 + (1-t)x_2) \\ &= tAx_1 + (1-t)Ax_2 \\ &= tf(x_1) + (1-t)f(x_2) \end{aligned} \tag{1}$$

  - Therefore, the linear function satisfies the convex inequality

2. Affine functions are convex

- $f(x) = Ax + b$ is convex where b is some vector in $\mathbb{R}$
- An Affine transformation is a linear transformation Ax plus translation b
  - All linear functions are affine functions but not vice versa
- **proof.**
  - A function f is convex if for $\forall x_1, x_2 \in \mathbb{R}; \forall t \in [0, 1]$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

  - By definition, an affine function is:

$$\begin{aligned} f(tx_1 + (1-t)x_2) &= A(tx_1 + (1-t)x_2) + b \\ &= tAx_1 + tb + (1-t)Ax_2 + (1-t)b \\ &= tf(x_1) + (1-t)f(x_2) \end{aligned} \tag{2}$$

  - Therefore, the affine function satisfies the convex inequality

3. Adding two convex functions results in a convex function

- $f(x) = h(x) + g(x)$ is a convex function
  - if $h(x)$ and $g(x)$ are convex
- **proof.**
  - A function f is convex if for $\forall x_1, x_2 \in \mathbb{R}; \forall t \in [0, 1]$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

  - Adding two convex functions:

$$\begin{aligned} f(tx_1 + (1-t)x_2) &= h(tx_1 + (1-t)x_2) + g(tx_1 + (1-t)x_2) \\ &\leq th(x_1) + tg(x_1) + (1-t)h(x_2) + \\ &(1-t)g(x_2) \\ &= tf(x_1) + (1-t)f(x_2) \end{aligned} \tag{3}$$

4. Composition with an affine mapping

- $f(x) = g(Ax + b)$ is convex if g is convex
- **proof.**

$$\begin{aligned} f(tx_1 + (1-t)x_2) &= g(A(tx_1 + (1-t)x_2) + b) \\ &= g(t(Ax_1 + b) + (1-t)(Ax_2 + b)) \\ &\leq tg(Ax_1 + b) + (1-t)g(Ax_2 + b) \\ &= tf(x_1) + (1-t)f(x_2) \end{aligned} \tag{4}$$

- Therefore, knowing that $Ax + b$ is convex it is sufficient to show that $f(z)$ is convex by replacing $Ax + b$ with z.
  - might be helpful in the assignment.

5. Pointwise maximum

- The max of two convex functions is convex
- $f = \max(f_1, f_2)$ is convex
- **proof.**

$$\begin{aligned} f(tx_1 + (1-t)x_2) &= \max(f_1(tx_1 + (1-t)x_2), f_2(tx_1 + (1-t)x_2)) \\ &\leq \max(tf_1(x_1) + (1-t)f_1(x_2), tf_2(x_1) + (1-t)f_2(x_2)) \\ &\leq \max(tf_1(x_1), tf_2(x_1)) + \\ &\max((1-t)f_1(x_2), (1-t)f_2(x_2)) \\ &= tf(x_1) + (1-t)f(x_2) \end{aligned} \tag{5}$$

5. Norms are convex functions

- For all norms $||x||_p = (\sum_{i=1}^{d} |x_i|^p)^{\frac{1}{p}}$ where $p \geq 1$ the following properties hold:
  - $||x|| \geq 0, \forall x \in R^d$
  - $||x|| = 0$ iff $x = 0$
  - $||ax|| = |a|||x||, \forall a \in R, x \in R^d$ (Homogeniety)
  - $||x_1 + x_2|| \leq ||x_1|| + ||x_2||, \forall x_1, x_2 \in R^d$ (Triangle inequality)
- **proof.** Norm functions are convex:

$$\begin{aligned} ||tx_1 + (1-t)x_2|| &\leq ||tx_1|| + ||(1-t)x_2|| \quad \text{(Triangle Inequality)} \\ &= t||x_1|| + (1-t)||x_2|| \quad \text{(Homogeniety)} \end{aligned} \tag{6}$$

6. Second-derivative test

- If the second derivative of a function $f(x)$ is positive $\forall x \in \mathbb{R}$ then f is convex
- **proof.**
- Using second order Taylor expansion, for some $\forall x_1, x_2 \in \mathbb{R}, \forall t \in [0, 1]$:

$$f(x_2) = f(x_1) + \nabla f(x_1)^T (x_2 - x_1) + (x_2 - x_1)^T \nabla^2 f(x_1 + t(x_2 - x_1))(x_2 - x_1) \tag{7}$$

- Since $\nabla^2 f(x) > 0$

$$(x_2 - x_1)^T \nabla^2 f(x_1 + t(x_2 - x_1))(x_2 - x_1) \geq 0 \tag{8}$$

- Therefore,

$$f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1) \tag{9}$$

6. Second-derivative test proof

- Let $x_1 < x_2$ and $y = tx_1 + (1-t)x_2$, then

$$\begin{aligned} f(x_1) &\geq f(y) + \nabla f(y)^T (y - x_1) \\ f(x_2) &\geq f(y) + \nabla f(y)^T (y - x_2) \end{aligned} \tag{10}$$

- Multiply the first inequality by t and second by $(1-t)$ and add them to get,

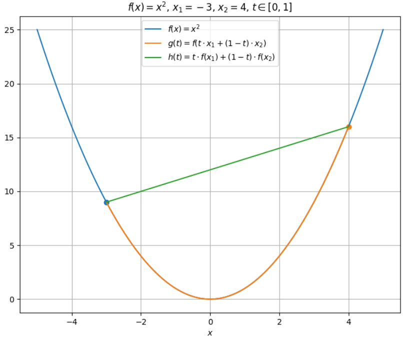
$$\begin{aligned} tf(x_1)+&(1-t)f(x_2) \geq tf(y) + (1-t)f(y) + \\ &t\nabla f(y)^T (y - x_1) + (1-t)\nabla f(y)^T (y - x_2) \\ &\Rightarrow tf(x_1) + (1-t)f(x_2) \geq f(y) + \\ &\nabla f(y)^T ((t-1)x_1 + (1-t)x_2) + \nabla f(y)^T ((t-1)x_2 + (1-t)x_1) \end{aligned} \tag{11}$$

- Therefore,

$$tf(x_1) + (1-t)f(x_2) \geq f(tx_1 + (1-t)x_2) \tag{12}$$

6. Second-derivative test

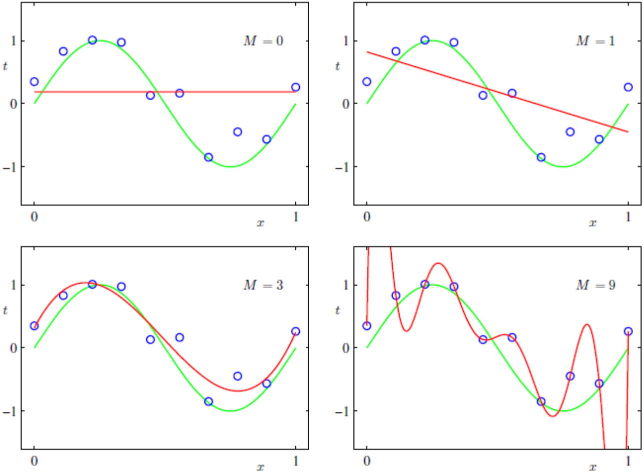
- Geometrically:
  - When $\nabla f(x)$ is negative, $f(x)$ decreases as x increases.
  - When $\nabla f(x)$ is positive, $f(x)$ increases as x increases.
  - Therefore, the minimum is at $x = a$ where the gradient switches sign.



Warning: Products

- The product of two convex functions is not necessarily convex.
- Consider $f(x) = x$ and $g(x) = -x$.
- Is $f(x)g(x) = -x^2$ a convex function?
  - Can you explain why?

Overfitting



Overfitting and Regularization

- Overfitting on the training set is a common problem and leads to worse test error.
- Models that are too "flexible" or "complex" for the available data will overfit.
  - Intuitively, the model is learning from spurious noise in the training set.
- Regularization tries to restrict the set of learnable models by adding a penalty to the loss function.

L2 Regularization

- Add the L2 norm of w to the loss function to penalize model complexity.
- The Loss function becomes

$$f(w) = L(w, X, y) + \frac{\lambda}{2} * ||w||_2^2$$

- $||w||_2^2$ will be large when the entries of w are large.
  - How does this penalize complex models?

L1: A Different Flavor

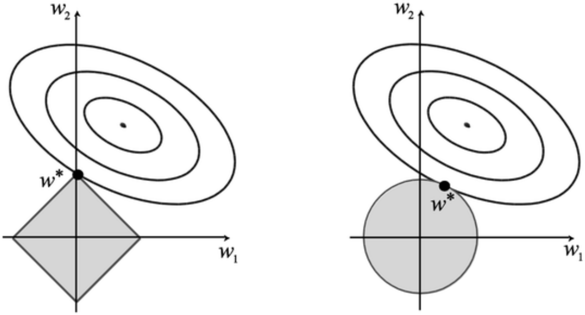
- Penalize with the L1 norm w instead of the L2 norm.
- The Loss function becomes

$$f(w) = L(w, X, y) + \lambda * ||w||_1$$

- How does this differ from L2 regularization?
  - Hint: Differentiability.
  - Hint: Size of penalties.

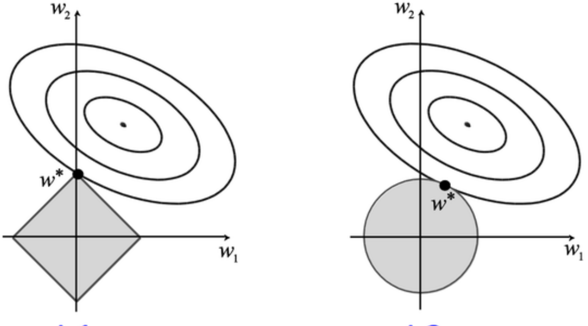
The Geometry Behind Regularization

- We can view L2 regularization as constraining $||w||_2^2$ to be less than some radius r.
  - r is uniquely determined by the choice of λ.
- Geometrically, we are restricting w to be in a hypersphere of radius r around the origin.
- Similarly, we can view L1 regularization as restricting w to be in a hypercube of side length r.



L1 vs L2 Regularization: Feature Selection

- L2 regularization does not perform feature selection.
  - Generally, elements of w are only set to zero as λ approaches infinity.
- L1 regularized regression does feature selection.
  - Elements of w can be set exactly to 0.
- The geometric interpretation of regularization gives useful intuition.



L1 vs L2 Regularization: Unique Solutions

- L2 regularized regression always has a unique solution.
- Why is this true? Think about the case where two features of X are identical.
  - Uniqueness is a common motivation for L2 regularization in statistics.
- L1 regularized regression does not always have a unique solution.
  - Try considering the case above again.

Questions on Regularization?

- Ask away!

Bonus Slide: Bias vs. Variance!

- Regularization is good for models with high sampling variance.
  - High Sampling Variance means the model parameters fluctuate significantly with different training sets.
- Regularization limits space of learnable models, which reduces variance.
- However, it introduces bias - the learned model isn't the "best" possible according to the training error.