On Parameter Estimation of Regression Models

CS189/289A: Introduction to Machine Learning

Stella Yu

UC Berkeley

5 September 2017

1 Review

1. Least squares

$$y = Ax \quad \Rightarrow \quad \min_x \|y - Ax\|^2 \quad \Rightarrow x = (A'A)^{-1}A'y \tag{1}$$

2. Ridge regression

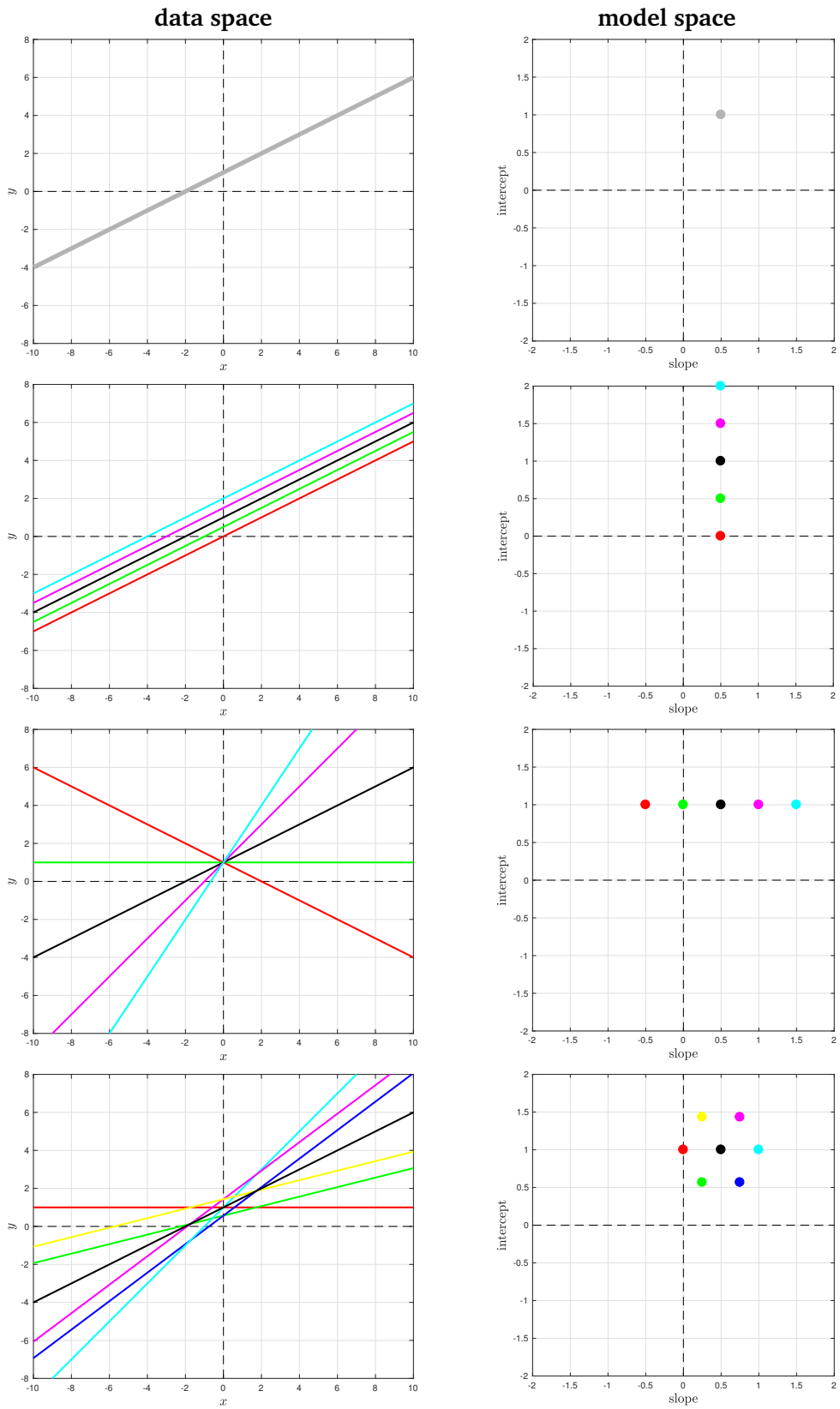
$$\min_x \|y - Ax\|^2 + \lambda\|x\|^2 \quad \Rightarrow x = (A'A + \lambda I)^{-1}A'y \tag{2}$$

3. Features
   A from polynomial terms of raw data; polynomial features as a universal feature representation; the importance of normalizing and conditioning the feature matrix; the model is still linear.
4. Cross-validation
   training, validation, test sets for selecting parameters, hypermeters, and algorithms/methods respectively.
5. Today
   Statistical, probabilistic views of regression model parameter estimation

2 The Underlying Model in the Data Space and Its Parameter Space

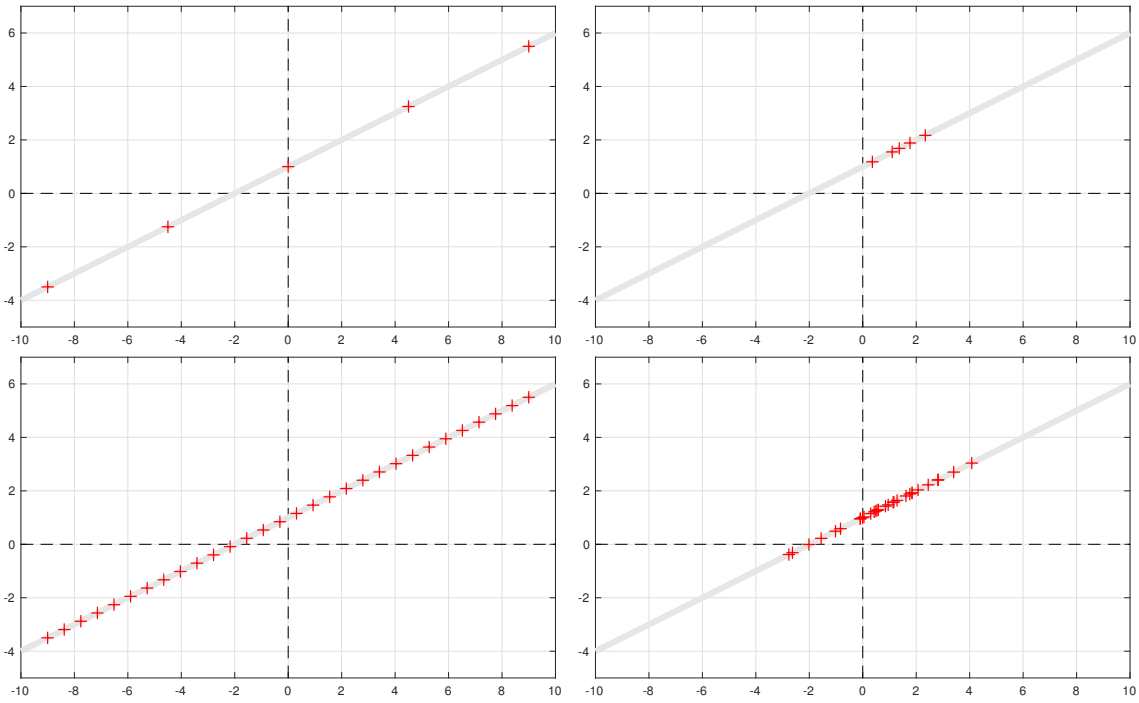
$$y = f(x) = \text{slope} \cdot x + \text{intercept} \tag{3}$$

- The underlying model of data points (x, y) is a line in the data space.
- The underlying model of data points (x, y) is a point in the model parameter space (slope,intercept).
- Learning the optimal model of data points (x, y) is to fit the points with a line in the data space, or it is equivalent to locate the optimal parameter point in the model space.


data space
model space
y
x
intercept
slope

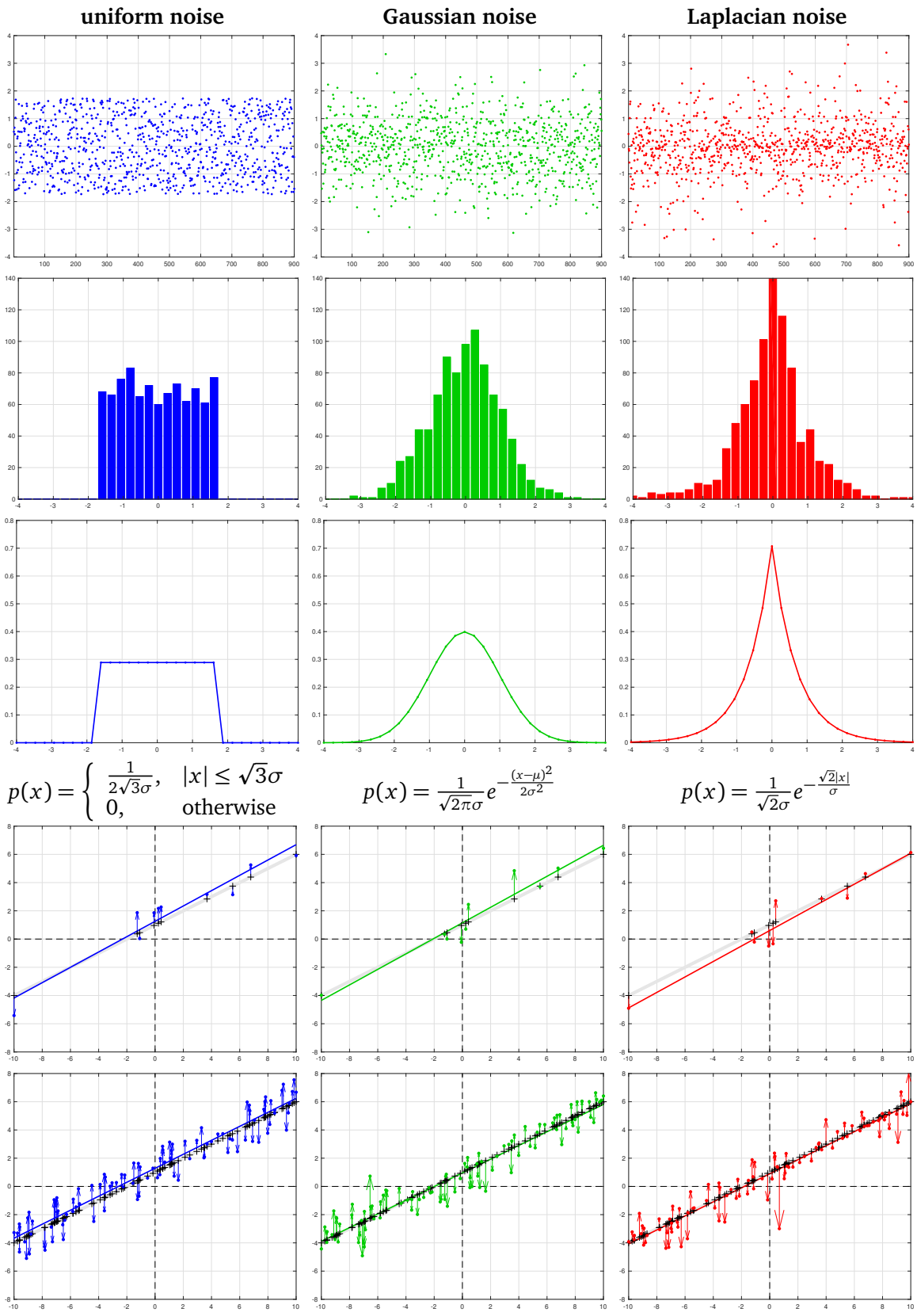
3 Data As Samples of the Underlying Model

Observed data are random samples with different distributions, coverages, and densities.



4 Data Generated with Additive White Noise

additive:	$Y = f(x) + N(x)$	(4)
random variable :	$N(x) \sim p(n(x))$	(5)
zero mean:	$E[N(x)] = 0, \quad \forall x$	(6)
independent & identically distributed:	$p(n(x)) = p(n), \quad \forall x$	(7)


uniform noise
Gaussian noise
Laplacian noise
$p(x) = \begin{cases} \frac{1}{2\sqrt{3}\sigma}, & |x| \le \sqrt{3}\sigma \\ 0, & \text{otherwise} \end{cases}$
$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
$p(x) = \frac{1}{\sqrt{2}\sigma} e^{-\frac{\sqrt{2}|x|}{\sigma}}$

5 Univariate Gaussian Distribution

random variable: $$X \sim p(x) \tag{8}$$

probability distribution: $$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{9}$$

mean: $$E[X] = \int_{-\infty}^{+\infty} x p(x) dx = \mu \tag{10}$$

variance: $$V[X] = E[(X-\mu)^2] = \int_{-\infty}^{+\infty} (x-\mu)^2 p(x) dx = \sigma^2 \tag{11}$$

parameters: $$X \sim \mathcal{N}(\mu, \sigma^2) \tag{12}$$

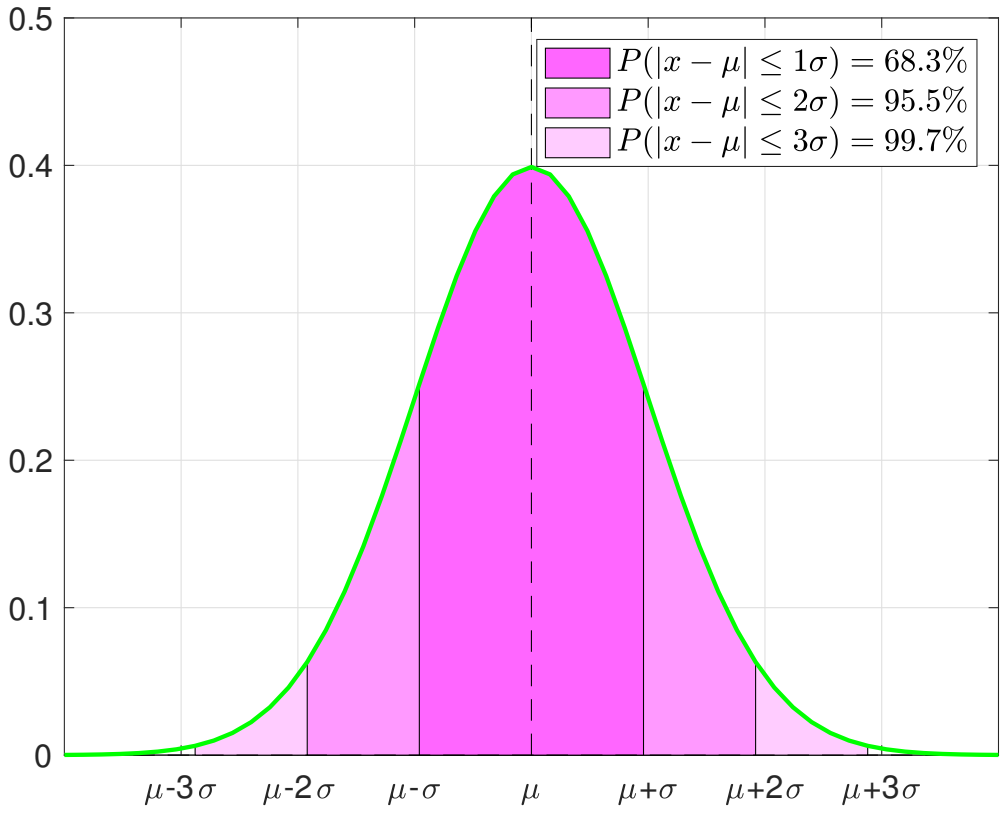
log-likelihood: $$\log P(X = x) = -\frac{(x-\mu)^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma) \tag{13}$$

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2), \qquad Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2) \tag{14}$$

linear combinations: $$E[aX + bY] = a\mu_X + b\mu_Y \tag{15}$$

if X, Y are independent: $$V[aX + bY] = a^2\sigma_X^2 + b^2\sigma_Y^2 \tag{16}$$

$$aX + bY \sim \mathcal{N}(a\mu_x + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2) \tag{17}$$



6 Maximum Likelihood Model Estimation

6.1 A Probabilistic Look at the Model Estimation Problem

- Given points of X sampled from an unknown probability distribution
- We observe Y as the sum of an unknown, non-random function f and random white noise:

$$Y = f(X) + N, \tag{18}$$

where f represents a consistent mapping relationship between X and Y.

- Once a sample of X is picked at x, the observed Y is a deterministic value $f(x)$ corrupted by some additive white noise N.

$$Y|(X = x, f) \quad = f(x) + N \sim \text{Noise-Model}(\text{mean } = f(x)). \tag{19}$$

- To find function h that estimates f, ideally, we just need to get rid of the random noise:

$$h(x) = E_Y[Y|X = x] = E_Y[f(X) + N|X = x] = f(x) + E[N] = f(x) \tag{20}$$

- If we can have multiple observations of Y, y_i, at the same point $X = x$, we could just compute the expectation of these y_i's, but that rarely happens, nor does it help us discover the common underlying model $f(x)$ linking y_i's across x_i's.
- The basic idea for estimating the underlying model given observed data is to score any model using the data, and pick the highest scoring point in the model space. Model estimation then becomes an optimization problem in the model parameter space.

6.2 Maximum Likelihood As a Model Scorer

- The log-likelihood of all m observations $\{(x_i, y_i), i = 1, \ldots, m\}$ is:

$$\log P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_m = y_m | X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m; f) \tag{21}$$
$$= \log P(y_1|x_1; f) \cdot P(y_2|x_2; f) \cdot \ldots \cdot P(y_m|x_m; f) \tag{22}$$
$$= \log P(y_1|x_1; f) + \log P(y_2|x_2; f) + \ldots + \log P(y_m|x_m; f) \tag{23}$$
$$= \sum_{i=1}^{m} \log P(y_i|x_i; f) \tag{24}$$

The larger value the $\log P(y_1, y_2, \ldots, y_m|x_1, x_2, \ldots, x_m; f)$, the more likely the observations (x_i, y_i) are generated by the model f.

- The maximum log-likelihood estimation (MLE) of model f is:

$$h = \arg\max_f \log P(y_1, y_2, \ldots, y_m|x_1, x_2, \ldots, x_m; f) \tag{25}$$
$$= \arg\max_f \sum_{i=1}^{m} \log P(y_i|x_i; f) \tag{26}$$

6.3 Maximum Likelihood Estimate Under the Gaussian Noise Model

The maximum likelihood estimation of f under the Gaussian noise assumption leads to the ordinary least square regression formulation.

- Assume that the noise is Gaussian with a fixed but unknown σ:

$$Y = f(X) + N, \quad N \sim \mathcal{N}(0, \sigma^2) \tag{27}$$

$$Y|X = x \sim \mathcal{N}(f(x), \sigma^2). \tag{28}$$

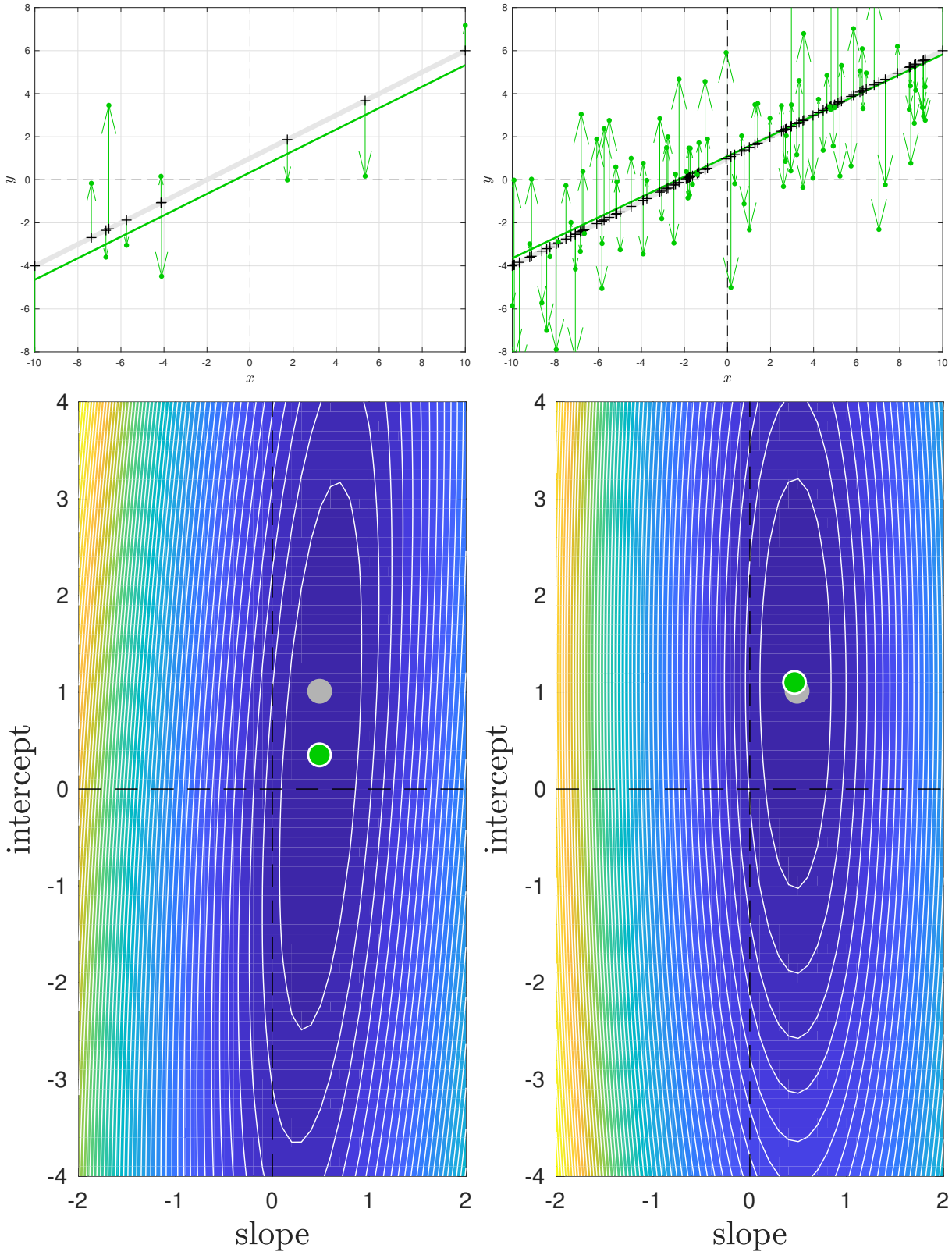
- The maximum log-likelihood estimation of f is:

$$h = \arg\max_f \log P(y_1, y_2, \ldots, y_m | x_1, x_2, \ldots, x_m; f) \tag{29}$$

$$= \arg\max_f \sum_{i=1}^{m} \log P(y_i | x_i; f) \tag{30}$$

$$= \arg\min_f \sum_{i=1}^{m} \frac{(y_i - f(x_i))^2}{2\sigma^2} + m \log(\sqrt{2\pi}\sigma) \tag{31}$$

$$= \arg\min_f \sum_{i=1}^{m} (y_i - f(x_i))^2 \tag{32}$$


y
x
intercept
slope

7 Bayesian Model Estimation

Whereas in maximum likelihood methods, we view the underlying model f to be fixed, in Bayesian learning, we consider the model f to be a random variable, and training data allow us to convert a distribution on this variable into a posterior probability density.

- The posterior probability of f given observations of (X, Y) and prior model distribution $P(f)$ is:

$$P(f|Y,X) = \frac{P(Y|X,f)P(f|X)}{\int_f P(Y|X,f)P(f|X)df} \tag{33}$$

$$= \frac{P(Y|X,f)P(f)}{\int_f P(Y|X,f)P(f)df}. \tag{34}$$

- The log-posterior probability of f is:

$$\log P(f|Y,X) = \log \frac{P(Y|X,f)P(f)}{\int_f P(Y|X,f)P(f)df} \tag{35}$$

$$= \log P(Y|X,f)P(f) - \underbrace{\log \int_f P(Y|X,f)P(f)df}_{\text{constant}} \tag{36}$$

$$= \underbrace{\log P(Y|X,f)}_{\text{data term}} + \underbrace{\log P(f)}_{\text{prior term}} - \text{constant} \tag{37}$$

  The better the f explaining the data and conforming to our prior knowledge of the model, the larger the posterior probability of f.

- The maximum a posterior estimation (MAP) of the model f:

$$h = \arg\max_f \log P(f|Y,X) \tag{38}$$

$$= \arg\max \log P(Y|X,f) + \log P(f) \tag{39}$$

- For Gaussian noise and prior models, given data $(x_i, y_i), i = 1, \ldots, m$, we have:

$$f \sim \mathcal{N}(f_0, \sigma_f^2 I) \tag{40}$$

$$Y|(X = x, f) \sim \mathcal{N}(f(x), \sigma^2) \tag{41}$$

$$h = \arg\max_f \left(-\frac{\sum_{i=1}^m (y_i - f(x_i))^2}{2\sigma^2} \right) + \left(-\frac{\|f - f_0\|^2}{2\sigma_f^2} \right) \tag{42}$$

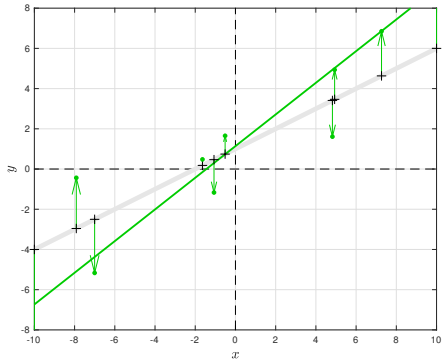
$$= \arg\min_f \underbrace{\sum_{i=1}^m (y_i - f(x_i))^2}_{\text{data term}} + \underbrace{\frac{\sigma^2}{\sigma_f^2}}_{\text{noise to signal ratio}} \cdot \underbrace{\|f - f_0\|^2}_{\text{prior term}} \tag{43}$$

  The larger the noise, the heavier weight towards the prior term.

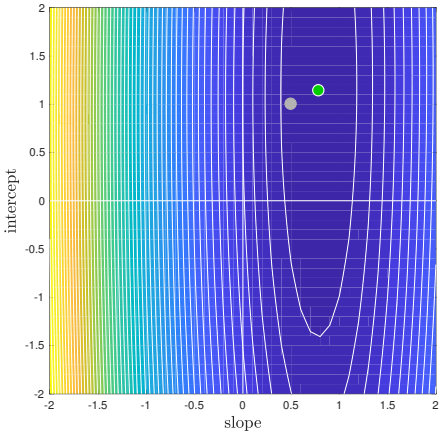
- When $f_0 = 0$, the MAP estimation becomes the ridge regression.

- When $\sigma_f \to \infty$, i.e., no model preference *a priori*, or
  when $\sigma = 0$, i.e., the training samples are noiseless,
  the noise to signal ratio $\frac{\sigma^2}{\sigma_f^2} = 0$,
  then the MAP estimation becomes the MLE estimation, i.e.,
  MLE is MAP with a uniform prior or noiseless training data.
- When $m \to \infty$, i.e., when the number of training samples increases, the prior term becomes less and less important. Data is knowledge.

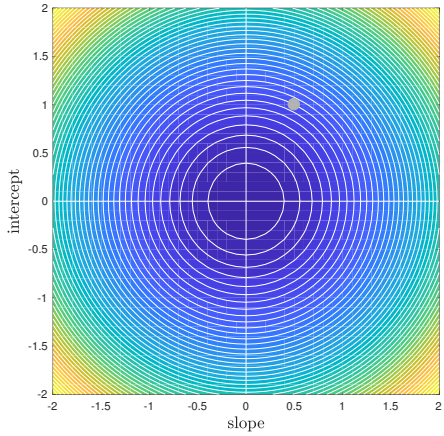
data



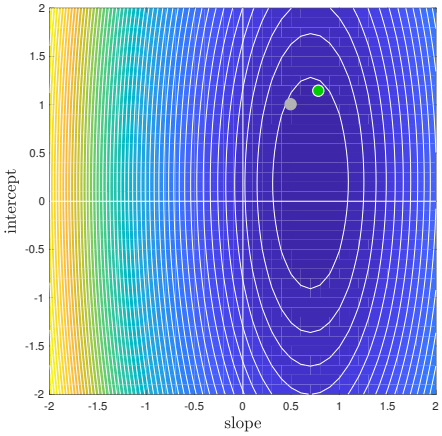
MLE



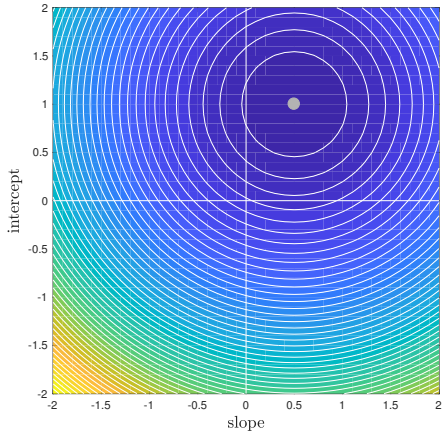
ridge prior



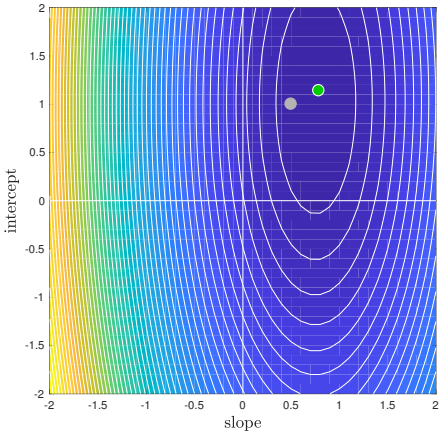
ridge MAP



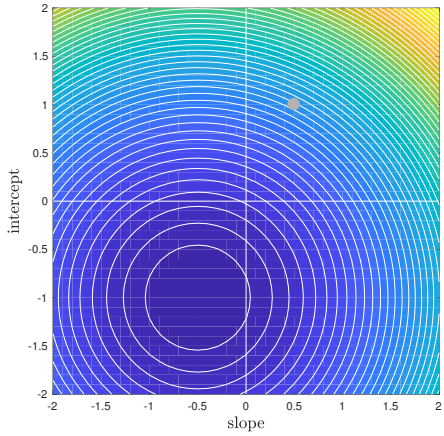
good prior



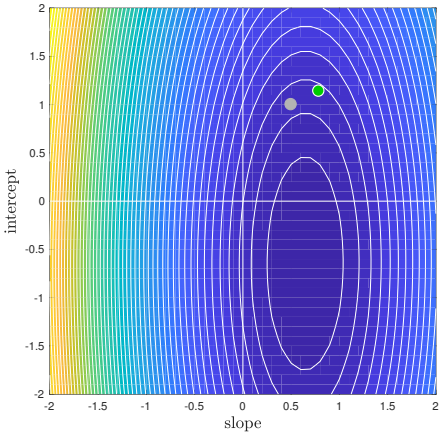
good MAP



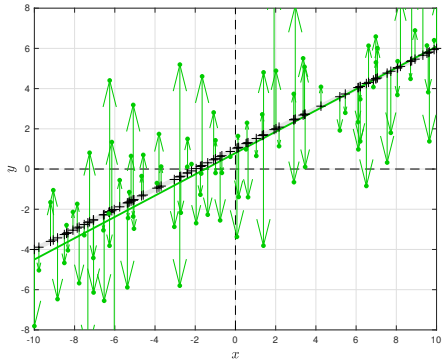
bad prior



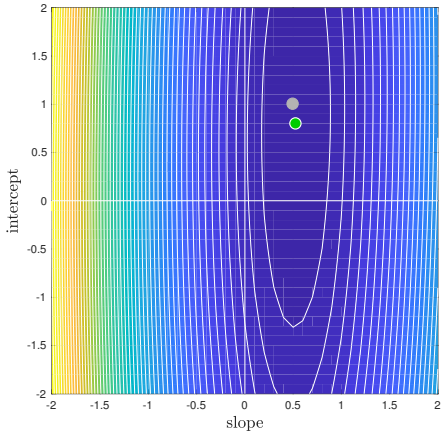
bad MAP



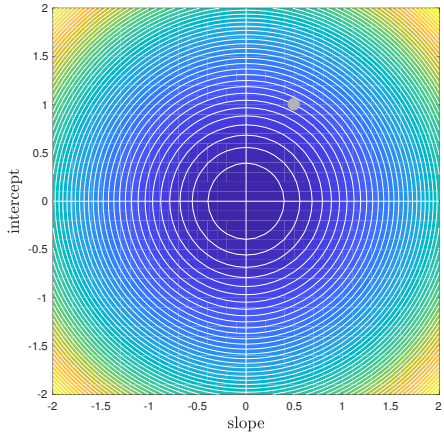
data



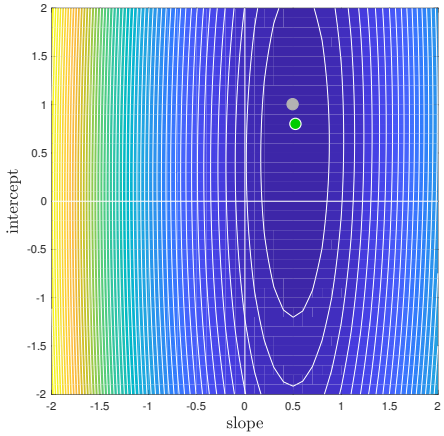
MLE



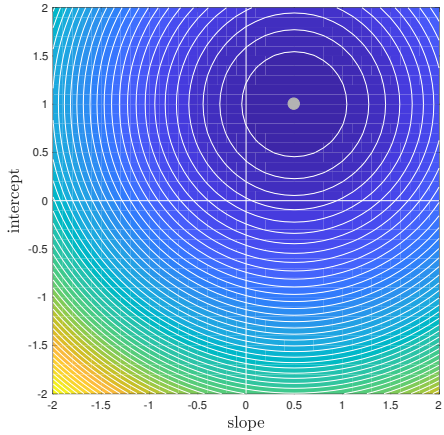
ridge prior



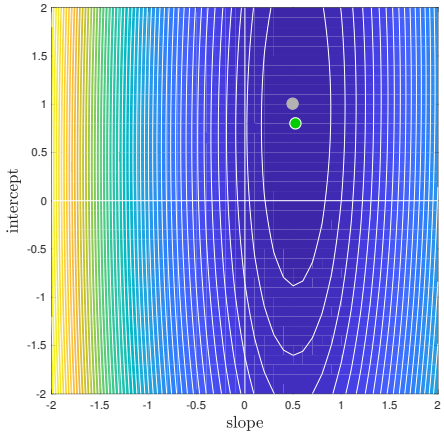
ridge MAP



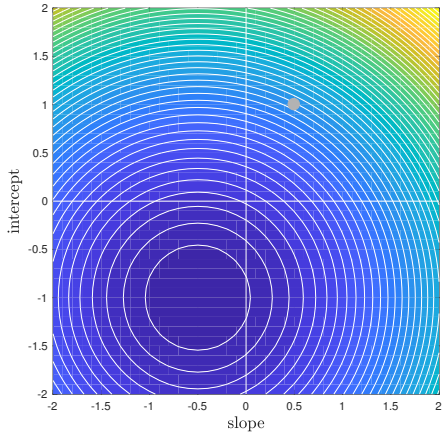
good prior



good MAP



bad prior



bad MAP