STAT 33612 Autumn 2025 Scribe: Pui Kuen Leung
November 11, 2025 Lecturer: Frederic Koehler
These notes have not received the scrutiny of publication. They could be missing important references, etc.

CGMT Applications, Motivating Example for Belief Propagation

1 CGMT applications

1.1 Re-deriving the classical asymptotic training error of OLS

Consider the linear model

$$Y = Xw^* + \xi$$

where $X \in \mathbb{R}^{n\times p}$ is a random matrix with i.i.d. $\mathcal{N}(0,1)$ entries and $\xi \sim \mathcal{N}(0,\sigma^2 I_n)$ is Gaussian noise. The ordinary least squares (OLS) estimator is

$$\widehat{w}_{\mathsf{OLS}} = \arg\min_{w\in\mathbb{R}^p} \|Y - Xw\|_2^2.$$

We work in the proportional asymptotics regime $p, n \to \infty$ and $\gamma = p/n \in (0,1)$. Recall from a previous lecture the classical result for the training error

$$\frac{1}{n}\|Y - X\widehat{w}_{\mathsf{OLS}}\|_2^2 \approx \sigma^2(1-\gamma).$$

We can re-derive this result using the Convex Gaussian Min-Max Theorem (CGMT). For convenience, we consider the equivalent optimization problem without the square:

$$\min_{w\in\mathbb{R}^p} \|Y - Xw\|_2$$

which has the same optimizer $\widehat{w}_{\mathsf{OLS}}$. Using the variational representation $\|v\|_2 = \max_{\|\lambda\|_2\le 1}\langle v,\lambda\rangle$, we can write this as a min-max problem:

$$\min_w \|Y - Xw\|_2 = \min_w \max_{\|\lambda\|_2\le 1}\langle Y - Xw, \lambda\rangle = \min_w \max_{\|\lambda\|_2\le 1}\langle X(w^*-w)+\xi, \lambda\rangle.$$

Take this to be the primary optimization (PO) problem. By CGMT, with high probability we have

$$\text{PO} \approx \text{AO},$$

where the associated auxiliary optimization problem is given by

$$\begin{aligned}\text{AO} &= \min_{w\in\mathbb{R}^p}\max_{\|\lambda\|_2\le 1}\left\{\langle \xi,\lambda\rangle + \langle g,\lambda\rangle\|w^*-w\|_2 + \langle h, w^*-w\rangle\|\lambda\|_2\right\}\\ &= \min_{w\in\mathbb{R}^p}\max_{\|\lambda\|_2\le 1}\left\{\langle \lambda, g\|w^*-w\|_2 + \xi\rangle + \langle w^*-w, h\rangle\|\lambda\|_2\right\},\end{aligned}$$

with $g \sim \mathcal{N}(0, I_n)$ and $h \sim \mathcal{N}(0, I_p)$ independent. Write $\lambda = tu$ where $t \in [0,1]$ and $\|u\|_2 = 1$. Letting $r := \|w^*-w\|_2$, the inner maximization problem becomes

$$\max_{t\in[0,1]}\max_{\|u\|_2=1} t\left(\langle gr+\xi, u\rangle + \langle w^*-w, h\rangle\right).$$

For a fixed t, the maximizer over u is given by the unit vector in the direction of $gr+\xi$, so the maximization becomes

$$\max_{t\in[0,1]} t(\|gr+\xi\|_2 + \langle w^*-w,h\rangle) = \begin{cases} 0 & \text{if } \|gr+\xi\|_2 + \langle w^*-w,h\rangle \le 0\\ \|gr+\xi\|_2 + \langle w^*-w,h\rangle & \text{if } \|gr+\xi\|_2 + \langle w^*-w,h\rangle > 0.\end{cases}$$

Thus, we can rewrite AO as

$$\text{AO} = \min_{w \in \mathbb{R}^p} \max\{0, \|g\|w^* - w\|_2 + \xi\|_2 + \langle w^* - w, h \rangle\},$$

Let us focus on the non-zero case

$$\min_{w \in \mathbb{R}^p} \{\|g\|w^* - w\|_2 + \xi\|_2 + \langle w^* - w, h \rangle\} = \min_{r \geq 0} \{\|gr + \xi\|_2 - r\|h\|_2\}$$

where we used the observation that the minimum occurs when we take w so that $w^* - w$ is in the negative direction of h. We now invoke standard facts about high-dimensional Gaussian vectors to see that:

- $\|h\|_2 \approx \sqrt{p}$, $\|g\|_2 \approx \sqrt{n}$, $\|\xi\|_2 \approx \sigma\sqrt{n}$
- g, ξ are independently drawn Gaussian random vectors in $\mathbb{R}^n$, so they are asymptotically orthogonal as $n \to \infty$, i.e.
$$\frac{|\langle g, \xi \rangle|}{\|g\|_2 \|\xi\|_2} \to 0 \quad \text{as } n \to \infty.$$

Thus,

$$\|gr + \xi\|_2^2 = r^2\|g\|_2^2 + \|\xi\|_2^2 + 2r\langle g, \xi \rangle \approx r^2 n + \sigma^2 n,$$

so

$$\min_{r \geq 0} \{\|gr + \xi\|_2 - r\|h\|_2\} \approx \min_{r \geq 0} \left\{\sqrt{r^2 n + \sigma^2 n} - r\sqrt{p}\right\}.$$

Notice that in our regime where $p/n = \gamma < 1$, $\sqrt{r^2 n + \sigma^2 n} - r\sqrt{p}$ is always strictly positive, so this rules out the case where $\lambda = 0$ is the optimizer for the inner maximization problem and we can safely write

$$\text{AO} = \min_{r \geq 0} \{\|gr + \xi\|_2 - r\|h\|_2\} \approx \min_{r \geq 0} \left\{\sqrt{r^2 n + \sigma^2 n} - r\sqrt{p}\right\}$$

To find the minimizer of the problem on the right-hand side, we differentiate the objective:

$$\frac{d}{dr} \left\{\sqrt{r^2 n + \sigma^2 n} - r\sqrt{p}\right\} = \frac{rn}{\sqrt{r^2 n + \sigma^2 n}} - \sqrt{p}.$$

Setting to 0 gives the stationary point

$$\frac{r}{\sqrt{r^2 + \sigma^2}} - \sqrt{\gamma} = 0 \implies \frac{r^2}{r^2 + \sigma^2} = \gamma \implies (1 - \gamma)r^2 = \sigma^2\gamma \implies r = \sigma\sqrt{\frac{\gamma}{1 - \gamma}}.$$

Then

$$\min_{r \geq 0} \left\{\sqrt{r^2 n + \sigma^2 n} - r\sqrt{p}\right\} = \sigma\sqrt{n}\left[\sqrt{\frac{1}{1 - \gamma}} - \sqrt{\frac{\gamma}{1 - \gamma}}\sqrt{\gamma}\right] = \sigma\sqrt{n}\sqrt{1 - \gamma}.$$

Therefore,

$$\frac{1}{n}\|Y - X\widehat{w}_{\mathsf{OLS}}\|_2^2 = \left(\frac{1}{\sqrt{n}}\|Y - X\widehat{w}_{\mathsf{OLS}}\|_2\right)^2 \approx \sigma^2 (1 - \gamma)$$

which is exactly what we wanted to show.

1.2 Estimation error for OLS

Following a similar derivation, we can also find the asymptotic estimation error in the same regime. We want to show that the estimation error $\|\widehat{w}_{\mathsf{OLS}} - w^*\|_2$ satisfies $\|\widehat{w}_{\mathsf{OLS}} - w^*\|_2 \approx r$, where $r = \sigma\sqrt{\frac{\gamma}{1-\gamma}}$. It is equivalent to show that for any small $\varepsilon > 0$, with high probability we have

$$\|\widehat{w}_{\mathsf{OLS}} - w^*\|_2 \notin [0, r-\varepsilon] \cup [r+\varepsilon, \infty).$$

Define the regions $A = \{w : \|w - w^*\|_2 < r - \varepsilon\}$, $B = \{w : r - \varepsilon \leq \|w - w^*\|_2 \leq r + \varepsilon\}$, and $C = \{w : \|w - w^*\|_2 > r + \varepsilon\}$. It suffices to show that

$$\min_{w \in A \cup C} \|Y - Xw\|_2 \gg \min_{w \in \mathbb{R}^p} \|Y - Xw\|_2.$$

Note that $A \cup C$ is not convex, so the left-hand side is not a convex optimization problem which we can apply CGMT to. However, we only need a lower bound here, and the Gaussian Min-Max theorem (GMT) guarantees PO $\geq$ AO holds with high probability even if the domain is not convex. We can do a similar analysis as above to write

$$\min_{w \in A \cup C} \|Y - Xw\|_2$$

as a min-max problem then use GMT and basic facts about high-dimensional Gaussian random vectors to get

$$\min_{w \in A \cup C} \|Y - Xw\|_2 \geq \min_{\tilde{r} \geq 0,\, \tilde{r} \notin [r-\varepsilon, r+\varepsilon]} \left\{ \sqrt{\sigma^2 n + \tilde{r}^2 n} - \tilde{r}\sqrt{p} \right\}.$$

Then the claim is proved once we note that

$$\min_{\tilde{r} \geq 0,\, \tilde{r} \notin [r-\varepsilon, r+\varepsilon]} \left\{ \sqrt{\sigma^2 n + \tilde{r}^2 n} - \tilde{r}\sqrt{p} \right\} \gg \min_{r \geq 0} \left\{ \sqrt{r^2 n + \sigma^2 n} - r\sqrt{p} \right\} \approx \min_{w \in \mathbb{R}^p} \|Y - Xw\|_2.$$

1.3 Other similar applications of CGMT

We mention several other estimators to which CGMT can be applied in a similar fashion to:

1. LASSO: This is the estimator $\arg\min_{w \in \mathbb{R}^p} \|Y - Xw\|_2^2 + \lambda\|w\|_1$. It can be rewritten via the variational formula $a^2/2 = \max_b\{ab - b^2/2\}$ as a min-max problem in a form which CGMT can be applied to.

2. Ridge regression: This is the same as LASSO but uses ℓ_2 penalty instead.

3. Interpolation: This corresponds to the limit of ridge regression as $\lambda \to 0$.

2 Motivating example for belief propagation

We now study a motivating problem for the next topic on belief propagation/cavity method/message passing algorithms. These three terms will be used more or less interchangeably in what follows. Consider the linear model $Y = Xw^* + \xi$ where $X \in \mathbb{R}^{n \times p}$ is a design matrix, $\xi \sim \mathcal{N}(0, \sigma^2 I_n)$ is Gaussian noise, and $w^* \in \{\pm 1\}^p$. How do we estimate w^*? A natural approach is to compute the MLE

$$\widehat{w}_{\mathsf{MLE}} = \arg\min_{w \in \{\pm 1\}^p} \|Y - Xw\|_2^2.$$

This is a nonconvex optimization problem. One way to solve it is to use integer programming, but this becomes computationally intractable for large p. Belief propagation gives us an alternative path. Viewing the constraint $w^* \in \{\pm 1\}^p$ as a prior, suppose we place a uniform prior $w^* \sim \mathsf{Unif}(\{\pm 1\}^p)$. By Bayes' theorem, the posterior $p(w^*|X, Y)$ is given by

$$\begin{aligned} p(w^*|X,Y) &\propto p(Y|X,w^*)p(w^*|X) \\ &\propto p(Y|X,w^*) \\ &\propto \exp\left(-\frac{\|Y - Xw^*\|_2^2}{2\sigma^2}\right) \\ &= \exp\left(\frac{-\|y\|_2^2 - \|Xw^*\|_2^2 + 2\langle Y, Xw^*\rangle}{2\sigma^2}\right) \\ &\propto \exp\left(\frac{-(w^*)^T X^T X w^* + 2\langle Y, Xw^*\rangle}{2\sigma^2}\right) \end{aligned}$$

The expression inside the exponential is a quadratic polynomial in the entries of w^*, and we call such a model an Ising model (to be defined formally in the next lecture). The cavity method can be thought of as an alternative to the replica method used to solve statistical physics models such as these.