

1 Maximum Likelihood Estimator

Three properties of maximum likelihood estimator:

1. θ_{ML} is consistent. $\theta_{ML} \rightarrow \theta_0$ when $n \rightarrow \infty$.
2. θ_{ML} is asymptotically normal. $\sqrt{n}(\theta_{ML} - \theta_0) \sim \mathcal{N}(0, I_n(\theta_0))$ when $n \rightarrow \infty$ and $I_n(\theta_0)$ is the fisher information.
3. θ_{ML} is asymptotically efficient. θ_{ML} minimizes $\mathbb{E}(\theta - \theta_0)^2$ when $n \rightarrow \infty$ because the asymptotic variance equals the Rao-Cramer bound (MLE is asymptotically unbiased). Note: when n is finite, θ_{ML} is not necessarily efficient, e.g., Stein estimator is universally more efficient for single sample.

Rao-Cramer bound: for any unbiased estimator $\hat{\theta}$ of θ_0 , $\mathbb{E}(\hat{\theta} - \theta_0)^2 \geq 1/I_n(\theta_0)$, where $I_n(\theta) = -\mathbb{E}(\frac{\partial^2}{\partial \theta^2} \log f(X; \theta) \mid \theta) = \mathbb{E}(\frac{\partial}{\partial \theta} \log f(X; \theta) \mid \theta)^2$ is the fisher information.

Sketch of Proof: define $\Lambda = \frac{\partial \log P(X; \theta)}{\partial \theta}$. Cauchy-Schwarz says $\text{Cov}^2(\Lambda, \hat{\theta}) \leq \text{Var}(\Lambda) \text{Var}(\hat{\theta}) = \mathbb{E}(\Lambda^2) \text{Var}(\hat{\theta})$ because $\mathbb{E}\Lambda = 0$. Note that $\text{Cov}(\Lambda, \hat{\theta}) = \mathbb{E}(\Lambda \hat{\theta}) = \int_X \hat{\theta}(x) \frac{\partial}{\partial \theta} f(x; \theta) dx = \frac{\partial}{\partial \theta} \int_X \hat{\theta}(x) f(x; \theta) dx = \frac{\partial}{\partial \theta} \mathbb{E}\hat{\theta} = 1$. Therefore, $\text{Var}(\hat{\theta}) \geq 1/\mathbb{E}(\Lambda^2)$.

However, when the dimension of problem goes to infinity while keeping the data-dim ratio fixed, MLE is biased and the p -values are unreliable.

2 Regression

Bias-Variance trade-off

Let D be the training dataset and \hat{f} be the predictive function. $\mathbb{E}_D \mathbb{E}_{Y|X} (\hat{f}(X) - Y)^2 = \mathbb{E}_D \mathbb{E}_{Y|X} [(\hat{f}(X) - \mathbb{E}_{Y|X} Y)^2 + (\mathbb{E}_{Y|X} Y - Y)^2] = \mathbb{E}_D (\hat{f}(X) - \mathbb{E}(Y \mid X))^2 + \mathbb{E}_D (\mathbb{E}(Y \mid X) - Y)^2 = \mathbb{E}_D (\hat{f}(x) - \mathbb{E}_D \hat{f}(x))^2 + (\mathbb{E}_D \hat{f}(x) - \mathbb{E}(Y \mid X))^2 + \mathbb{E}_D (\mathbb{E}(Y \mid X) - Y)^2$. It means that expected square error (training) = variance of prediction + squared bias + variance of noise. The optimal trade-off is achieved by avoiding under-fitting (large bias) and over-fitting (large variance). Note that here the variance of output is computed by refitting the regressor on a new dataset.

Regularization

Ridge and Lasso can be viewed as MAP (maximum a posterior) estimation. A Gaussian prior on β is equivalent to Ridge and a Laplacian prior is equivalent to Lasso. Using

SVD, we get Ridge has built-in model selection: $X\beta^{\text{Ridge}} = \sum_{j=1}^d [d_j^2/(d_j^2 + \lambda)] u_j u_j^T Y$ (each $u_j u_j^T Y$ can be viewed as a model). Lasso has more sparse estimations because the gradient of regularization does not shrink as in the case of Ridge.

3 BLR and GP

Bayesian Linear Regression

Model $Y = X\beta + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Prior $\beta \sim \mathcal{N}(0, \Lambda^{-1})$. Posterior $\beta \mid X, Y \sim \mathcal{N}(\mu_\beta, \Sigma_\beta)$, where $\mu_\beta = (X^T X + \sigma^2 \Lambda)^{-1} X^T Y$ and $\Sigma_\beta = (\sigma^{-2} X^T X + \Lambda)^{-1}$.

Gaussian Process

$Y = \begin{pmatrix} Y_0 \\ Y_1 \end{pmatrix}$ is the combination of observed and prediction value. Assume a Gaussian prior of $\mathcal{N}(0, K + \sigma^2 I)$, where $K_{ij} = k(x_i, x_j)$ is kernel. GP regression is the conditional/Posterior distribution on Y_0 , $\mathbb{E}[Y_1 | Y_0] = K_{10}(\sigma^2 I_0 + K_{00})^{-1} Y_0$, $\text{Cov}[Y_1] = \sigma^2 I_1 + K_{11} - K_{10}(\sigma^2 I_0 + K_{00})^{-1} K_{01}$. Bayesian LR is a special case of GP with linear kernel $k(x, y) = x^T \Lambda^{-1} y$.

Kernel Function

A function is a kernel iff (1) symmetry $k(x, x') = k(x', x)$ and (2) semi-positive definite $\int_{\Omega} k(x, x') f(x) f(x') dx dx' \geq 0$ for any $f \in L_2$ and $\Omega \in \mathcal{R}^d$ (continuous) or $K(X) \geq 0$ (discrete). The latter is equivalent to (1) $a^T K a \geq 0, \forall a$ or (2) $k(x, x') = \phi(x)^T \phi(x')$ for some ϕ .

Kernel Construction

If $k_{1,2}$ are valid kernels, then followings are valid: (1) $k(x, x') = k_1(x, x') + k_2(x, x')$. (2) $k(x, x') = k_1(x, x') \cdot k_2(x, x')$. Proof: expand by Mercer's thm. (3) $k(x, x') = c k_1(x, x')$ for constant $c > 0$. (4) $k(x, x') = f(k_1(x, x'))$ if f is a polynomial with positive coefficients or the exp. Proof: polynomial can be proved by applying the product, positive scaling and addition. Exp can be proved by taking limit on the polynomial. (5) $k(x, x') = f(x) k_1(x, x') f(x')$. (6) $k(x, x') = k_1(\phi(x), \phi(x'))$ for any function ϕ .

Example: RBF kernel $k(x, y) = e^{-\|x-y\|^2/2\sigma^2} = e^{-\|x\|^2/2\sigma^2} \times e^{x^T y/2\sigma^2} \times e^{-\|y\|^2/2\sigma^2}$ is valid. (1) $x^T y$ linear kernel is valid (2) then $\exp(\frac{1}{\sigma^2} x^T y)$ is valid, (3) let $f(x) = \exp(-\frac{1}{2\sigma^2} \|x\|^2)$, by rules $f(x)k(x, y)f(y)$ RBF is valid.

Mercer's Theorem: Assume $k(x, x')$ is a valid kernel. Then there exists an orthogonal basis

e_i and $\lambda_i \geq 0$, s.t. $k(x, x') = \sum_i \lambda_i e_i(x) e_i(x')$.

4 Linear Methods for Classification

Concept Comparison

1. Probabilistic Generative, modeling $p(x, y)$: (1) can create new samples, (2) outlier detection, (3) probability for prediction, (4) high computational cost and (5) high bias.
2. Probabilistic Discriminative, modeling $p(y \mid x)$: (1) probability for prediction, (2) medium computational cost and (3) medium bias.
3. Discriminative, modeling $y = f(x)$: (1) no probability for prediction, (2) low computational cost and (3) low bias.

Infer $p(x, y)$ for classification problems

Use $p(x, y) = p(y)p(x \mid y)$. Since y has finite states, model $p(y)$ and $p(x \mid y)$ for different y . The modeling requires to (1) guess a distribution family and (2) infer parameters by MLE.

Compute $p(y \mid x)$ by discriminant analysis (DA) Linear DA

Goal: classify a sample into two Gaussian distribution with $\Sigma_0 = \Sigma_1$. After calculation, $p(y = 1 \mid x) = 1/(1 + \exp(-\log \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)})) = 1/(1 + \exp(w_1^T x + w_0))$ since the quadratic term is eliminated due to $\Sigma_0 = \Sigma_1$.

Quadratic DA

Goal: classify a sample into two Gaussian distribution with $\Sigma_0 \neq \Sigma_1$. After calculation, $p(y = 1 \mid x) = 1/(1 + \exp(x^T W x + w_1^T x + w_0))$.

Optimization Methods

Optimal Learning Rate for Gradient Descent

Goal: find $\eta^* = \text{argmin}_\eta L(w^k - \eta \cdot \nabla L(w^k))$.

By Taylor expansion of $L(w^{k+1})$ at w^k and solve for the optimal η , we get $\eta^* =$

$$\frac{\|\nabla L(w^k)\|^2}{\nabla L(w^k)^T H_L(w^k) \nabla L(w^k)}.$$

However, naive gradient descent has two weaknesses: (1) it often has a zig-zag behavior, especially in a very narrow, long and slightly downward valley; (2) the gradient update is small near the stationary point. This can be mitigated by adding a momentum term in the update: $w^{k+1} = w^k - \eta \nabla L(w^k) + \mu^k (w^k - w^{k-1})$ which speeds the update towards the "common" direction.

Newton's Method

Taylor-expand $L(w)$ at w_k to derive the optimal w^{k+1} : $L(w) \approx L(w^k) + (w - w^k)^T \nabla L(w^k) + \frac{1}{2}(w - w^k)^T H_L(w^k)(w - w^k) \Rightarrow w^{k+1} = w^k -$

$$H_L^{-1}(w^k) \nabla L(w^k).$$

Pros: (1) better updates compared to GD since it uses the second Taylor term and (2) does not require learning rate.

Cons: requires H_L^{-1} which is expensive.

Bayesian Method

In most cases, the posterior is intractable. Use approximation of posterior instead.

Laplacian Method

Idea: approximate posterior near the MAP estimation with a Gaussian distribution. $p(w \mid X, Y) \propto p(w, X, Y) \propto \exp(-R(w))$, where $R(w) = -\log p(w, X, Y)$. Let $w^* = \text{argmin} R(w)$ be the MAP estimation and Taylor-expand $R(w)$ at w^* : $R(w) \approx R(w^*) + \frac{1}{2}(w - w^*)^T H_R(w^*)(w - w^*)$. Therefore, $p(w \mid X, Y) \propto \exp(-R(w^*) - \frac{1}{2}(w - w^*)^T H_R(w^*)(w - w^*))$ and thus $(w \mid X, Y) \sim \mathcal{N}(w^*, H_R^{-1}(w^*))$.

AIC & BIC

- Define BIC = $k \log N - 2 \log \hat{L}$, where k is #parameters and \hat{L} is the likelihood $p(x \mid w^*)$. A lower BIC means a better model.

- Define AIC = $2k - 2 \log \hat{L}$. A lower AIC means a better model.

LDA by loss minimization

Perceptron

Goal: for $y_i \in \{0, 1\}$, find w , s.t. $y_i w^T x_i > 0$ for any i . The classification function is $c(x) = \text{sgn}(w^T x)$.

$L(y, c(x)) = 0$ if $y w^T x > 0$ and $L(y, c(x)) = -y w^T x$ o.w. By gradient descent, the Perceptron is guaranteed to converge if (1) the data is linearly separable, (2) learning rate $\eta(k) > 0$, (3) $\sum_k \eta(k) \rightarrow +\infty$ and (4) $(\sum_k \eta(k)^2)/(\sum_k \eta(k))^2 \rightarrow 0$. However, there exists multiple solutions if the data is linearly separable.

Fisher's LDA

Idea: project the two distribution into one dimension and maximize the ratio of the variance between the classes and the variance within the classes, i.e., $\max(w^T u_1 - w^T u_0)^2 / (w^T S w)$, where $S = \Sigma_0 + \Sigma_1$. Let gradient be zero and solve for w^* , we get $w^* \propto S^{-1}(u_1 - u_0)$.

We first compute w^* and fit distributions of the two-class projection. Then apply Bayesian decision theory to make classification.

5 Optimization with Constraint

Problem $\min_x f(x)$ s.t. $g_{i \in [I]}(x) \leq 0$ and $h_{j \in [J]}(x) = 0$. Solve it with **KKT Cond**: (1) Stationary $\nabla f + \sum_i \lambda_i \nabla g_i + \sum_j \mu_j \nabla h_j = 0$, (2)

$h_j(x) = 0$, (3) primal feasibility $g_i(x) \leq 0$, (4) dual feasibility $\lambda_i \geq 0$, (5) complementary slackness $\lambda_i g_i(x) = 0$.

Weak Duality: Lagrangian $L(x, \lambda, \mu) = f(x) + \lambda^\top g(x) + \mu^\top h(x)$, $\lambda > 0$. Dual function $F(\lambda, \mu) := \min_x L(x, \lambda, \mu)$. Denote \tilde{x} optima of original problem, then $\lambda^\top g(\tilde{x}) + \mu^\top h(\tilde{x}) \leq 0, \forall \lambda, \mu$, $F(\lambda, \mu) = \min_x L(x, \lambda, \mu) \leq L(\tilde{x}, \lambda, \mu) \leq f(\tilde{x}) = \min_{x, h(x)=0, g(x) \leq 0} f(x)$

Strong Duality in Convex Optimization

If Slater's cond (1) f convex (2) g convex (3) h linear (4) $\exists \bar{x}$ s.t. $g_i(\bar{x}) < 0$ and $h_j(\bar{x}) = 0$, then Strong Duality $\max_{\lambda, \mu} F(\lambda, \mu) = \min_{x, h(x)=0, g(x) \leq 0} f(x)$ holds.

6 Support Vector Machine

Linear Separable Case

Primal: $\max_{w,b} \left\{ \frac{1}{\|w\|} \min_i y_i (w^\top x_i + b) \right\} \Leftrightarrow \max_{w,b,t} t$ s.t. $\forall i, t \leq y_i (w^\top x_i + b)$ and $\|w\| = 1 \Leftrightarrow \min_{w,b} \frac{1}{2} w^2$ s.t. $\forall i, 1 \leq y_i (w^\top x_i + b)$
 (1) KKT cond: $\forall i, \alpha_i \geq 0, (1 - y_i (w^\top x_i + b)) \leq 0, \alpha_i (1 - y_i (w^\top x_i + b)) = 0$
 (2) **Dual:** $\max_{\alpha} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$ s.t. $(\alpha_i \geq 0) \wedge (\sum_i \alpha_i y_i = 0)$

Non-separable Case

Introduce slack variables $\xi_i := \max\{1 - y_i (w^\top x_i + b), 0\} = [1 - y_i (w^\top x_i + b)]_+$ into loss.
Primal: $\min_{w,b} \frac{1}{2} w^2 + C \sum_i \xi_i = \min_{w,b} \frac{1}{2} w^2 + C [1 - y_i (w^\top x_i + b)]_+$. Hinge loss $[1 - x]_+$.
 Equivalent form: $\min_{w,b} \frac{1}{2} w^2 + C \sum_i \xi_i$ s.t. $y_i (w^\top x_i + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$
Dual: $\max_{\alpha} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$ s.t. $\sum_i \alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C$

Multi-class SVM

$\min_{w=[w_{0:K-1}], b=[b_{0:K-1}]} \frac{1}{2} \|w\|^2 + \sum_i C \xi_i$ s.t. $\xi_i \geq 0$ and $(w_{y_i}^\top x + b_{y_i}) - (w_y^\top x + b_y) \geq 1 - \xi_i, \forall y \neq y_i$

Structural SVM

y is structured, e.g. trees, maximum margin between y_i, y_j depends on their similarity, so the condition changes to $w^\top \Psi(x_i, y_i) - w^\top \Psi(x_i, y_j) \geq \Delta(y_i, y_j) - \xi_i, \forall y \neq y_i$.

7 Ensemble

Bagging Each bagged estimator have bias $\beta = \mathbb{E}(y - b(x))^2$, variance $\sigma^2 = \text{Var} b(x)$ co-variance $\rho^2 = \text{Cov}(b(x), b'(x))/\sigma^2$. Then $\mathbb{E}(y - \sum_m b^{(m)}(x)/M)^2 = \beta^2 + \sum_m \mathbb{E}(\beta - b^{(m)}(x))^2/M^2 = \beta^2 + \sigma^2/M + \sigma^2 \rho^2 (1 - 1/M)$. In class we assume $\rho = 0$. Anyway Bagging reduces variance.

Random Forest is a case of Bagging. Bagging induces implicit regularization.

Adaboost Initial $w_i^{(0)} = 1/n$. For $t \in [M]$, (1) train $f_t(x) = \arg\min_{b(x)} \sum w_i^{(t)} \mathbb{I}_{\{y_i \neq b(x_i)\}}$ (2) error $\epsilon_t = (\sum w_i^{(t)} \mathbb{I}_{\{y_i \neq f_t(x_i)\}}) / \sum w_i^{(t)}$ (3) estimator weight $\alpha_t = \log(\frac{1-\epsilon_t}{\epsilon_t})$ (4) data weight $w_i^{(t+1)} = w_i^{(t)} e^{\alpha_t \mathbb{I}_{\{y_i \neq f_t(x_i)\}}}$
Prediction $\hat{c} = \text{sgn}(\sum_{t=1}^M \alpha_t f_t(x))$

Gradient Boosting Initial $f_0(x) = 0$. For $t \in [M]$, (1) train $(\alpha_t, b^{(t)}) \leftarrow \arg\min_{\alpha > 0, b \in \mathcal{H}} \sum_{i=1}^n L(y_i, \alpha b(x_i) + f_{t-1}(x_i))$ (2) update function $f_t(x) \leftarrow \alpha_t b^{(t)}(x) + f_{t-1}(x)$. **Prediction** $\hat{c}(x) = \text{sgn}(f_M(x))$. Adaboost is GB with $L(y, \hat{y}) = e^{-y\hat{y}}$.

8 Generative Models

ELBO $\log p(y) = \log \int p(y | \theta) p(\theta) d\theta = \log \mathbb{E}_{\theta \sim q} \left[p(y | \theta) \frac{p(\theta)}{q(\theta)} \right] \geq \mathbb{E}_{\theta \sim q} \left[\log \left(p(y | \theta) \frac{p(\theta)}{q(\theta)} \right) \right] = \mathbb{E}_{\theta \sim q} [\log p(y | \theta)] - KL(q \| p(\cdot))$

VAE Goal: Find a latent representation z of x with simple prior $p_\theta(z)$. Problem: $p_\theta(x) = \mathbb{E}_\theta p(x|z)$ intractable. Solution: use encoder net $q_e(x|z)$ and $q_d(z|x)$ to model conditional and posterior prob.

ELBO for VAE training loss $l = \sum \log(p_\theta(x_i))$

$$\begin{aligned} \log(p_\theta(x_i)) &= \mathbb{E}_{z \sim q_\phi(z|x_i)} [\log p_\theta(x_i)] = \mathbb{E}_z \left[\log \frac{p_\theta(x_i|z)p_\theta(z)}{p_\theta(z|x_i)} \right] \\ &= \mathbb{E}_z \left[\log \frac{p_\theta(x_i|z)p_\theta(z)}{p_\theta(z|x_i)} \frac{q_\phi(z|x_i)}{q_\phi(z|x_i)} \right] \\ &= \mathbb{E}_z [\log p_\theta(x_i|z)] - \mathbb{E}_z \left[\log \frac{q_\phi(z|x_i)}{p_\theta(z)} \right] + \mathbb{E}_z \left[\log \frac{q_\phi(z|x_i)}{p_\theta(z|x_i)} \right] \\ &= \underbrace{\mathbb{E}_z [\log p_\theta(x_i|z)] - D_{KL}(q_\phi(z|x_i) \| p_\theta(z))}_{\mathcal{L}(x_i, \theta, \phi)} + \underbrace{D_{KL}(q_\phi(z|x_i) \| p_\theta(z|x_i))}_{\geq 0} \end{aligned}$$

Generative Adversarial Network: Generator G and Discriminator D . Optimize $\min_G \max_D V(D, G)$ where $V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))]$

9 Convergence of SGD, Robbins-Monro

Loss gradient $\ell(\cdot)$, SGD update $z^{(t)} \leftarrow \ell(\theta^{(t)} + \gamma^{(t)}, \theta^{(t+1)} \leftarrow \theta^{(t)} - \eta^{(t)} z^{(t)}, \gamma^{(t)}$ noise.

Problem: Whether $\theta^\infty \rightarrow \arg_{\theta^*} \mathbb{E}[\ell(\theta^*)] \triangleq 0$?

Assume: (1) $\mathbb{E}[\gamma] = 0$, (2) $\mathbb{E}[\gamma^2] = \sigma$ (3) $(\theta - \theta^*) \ell(\theta) > 0, \forall \theta \neq \theta^*$ (4) $\exists b, \ell(\theta) < b, \forall \theta$. If (1) $\eta^{(t)} \rightarrow 0$ (2) $\sum_{t < \infty} \eta(t) = \infty$ (3) $\sum_{t < \infty} \eta^2(t) < \infty$, then $\mathbb{P}(\theta^* = \theta^{(t)}) \xrightarrow[t \rightarrow \infty]{} 1$.

Proof: $\mathbb{E}[(\theta^{(t+1)} - \theta^*)^2] = \mathbb{E}[(\theta^{(t)} - \theta^*) - \eta^{(t)} l(\theta^{(t)}) - \eta^{(t)} \gamma^{(t)}]^2$. $\gamma^{(t)}$ independent with $\theta^{(t)}, \ell(\theta^{(t)})$, so LHS = $\mathbb{E}[(\theta^* - \theta^{(t)})^2] - 2\eta^{(t)} \mathbb{E}[\ell(\theta^{(t)}) (\theta^* - \theta^{(t)})] + \eta^{(t)2} (\mathbb{E}[\ell^2(\theta^{(t)})] + \mathbb{E}[\gamma^2(t)]) \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{(i)})] + \sum_{i \leq t} \eta^2(i) (b^2 + \sigma^2)$ Since $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] \leq \dots \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta(i) \mathbb{E}[\ell(\theta^{(i)}) (\theta^* - \theta^{($

A **Appendix**

(1) $\partial_x(AB) = A\partial_x B + (\partial_x A)B$, (2) $\partial_x A^{-1} = -A^{-1}(\partial_x A)A^{-1}$,
(3) $\partial_x \ln \det A = \text{Tr}\left(A^{-1}\partial_x A\right)$,

Define $(\partial_A f)_{ij} := \partial_{a_{ji}} f$, then (4) $\partial_A \text{Tr}(BA) = \partial_A \text{Tr}(AB) =$
 B , (5) $\partial_A \ln \det A = A^{-1}$, (6) $\partial_A \text{Tr}(ABA^\top) = (B + B^\top)A^\top$,

$\mathcal{N}(\mu, \Sigma) = (2\pi)^{-d/2} |\Sigma|^{-1/2} e^{-(x-\mu)^T \Sigma^{-1} (x-\mu)/2}$, **Conditional**
 $\mathbb{E}[y_2|y_1] = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(y_1 - \mu_1)$, $\text{Cov}[y_2 \mid y_1] = \Sigma_{22} -$
 $\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. **Marginal** $\mathbb{E}(y_2) = \mu_2$, $\text{Cov}[y_2] = \Sigma_{22}$

$(A + UC^{-1}V)^{-1} = A^{-1} - A^{-1}U(C + VA^{-1}U)^{-1}VA^{-1}$.