Chapter 3
Risk Minimization and Rademacher Complexity II

Ernest K. Ryu
Seoul National University

Mathematical Machine Learning Theory
Spring 2024

Outline

Calibration

$\mathcal{R}_{\Phi^{0\text{-}1}} - \mathcal{R}^\star_{\Phi^{0\text{-}1}} \overset{?}{\leq} H(\mathcal{R}_\Phi - \mathcal{R}^\star_\Phi)$

Calibration function with convex losses

Ridge least squares regression

ℓ^2-regularized estimation

Binary classification

Consider the binary classification problem, where $\tilde{\mathcal{Y}} = \mathcal{Y} = \{-1, +1\}$ and $\ell(y', y) = \mathbf{1}_{\{y' \neq y\}}$. So

$$\mathcal{R}[f] = \mathop{\mathbb{E}}_{(X,Y)\sim P}[\ell(f(X), Y)] = \mathop{\mathbb{P}}_{(X,Y)\sim P}(f(X) \neq Y).$$

Define

$$\eta(X) = \mathbb{P}(Y = +1 \,|\, X).$$

Assume $\eta(X) \neq 1/2$ with probability 1. Then,

$$f^\star(X) = \begin{cases} -1 & \text{if } \eta(X) < 1/2 \\ +1 & \text{if } \eta(X) > 1/2 \end{cases}$$

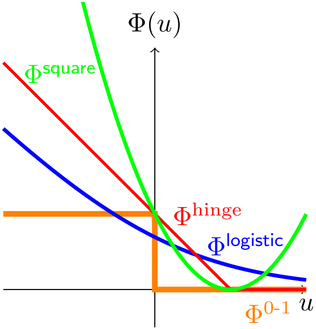
is a Bayes predictor, and

$$\mathcal{R}^\star = \mathop{\mathbb{E}}_{X\sim P_X}[\min\{1 - \eta(X), \eta(X)\}].$$

Surrogate loss

We replace $\Phi^{0\text{-}1}(u)$ with a surrogate loss such as

$$\Phi^{\mathsf{hinge}}(u) = \max\{1-u, 0\}$$
$$\Phi^{\mathsf{logistic}}(u) = \log(1+e^{-u})$$
$$\Phi^{\mathsf{square}}(u) = (1-u)^2,$$



which are nice continuous, convex functions, and solve the continuous convex optimization problem

$$\underset{g}{\text{minimize}} \quad \underbrace{\mathbb{E}_{(X,Y)\sim P}[\Phi(Yg(X))]}_{=\mathcal{R}_\Phi[g]}$$

or its approximation

$$\underset{g}{\text{minimize}} \quad \underbrace{\frac{1}{N}\sum_{i=1}^{N}\Phi(Y_i g(X_i)).}_{=\hat{\mathcal{R}}_\Phi[g]}$$

Binary classification with square loss

Consider the square surrogate loss

$$\mathcal{R}_{\Phi^{\mathsf{square}}}[g] = \mathop{\mathbb{E}}_{(X,Y)\sim P}[(1 - Yg(X))^2] = \mathop{\mathbb{E}}_{(X,Y)\sim P}[(g(X) - Y)^2].$$

Bayes predictor has a simple analytic form:

$$\begin{aligned} g^\star(X) &= \mathbb{E}[Y \mid X] = -1 \cdot \mathbb{P}(Y = -1 \mid X) + 1 \cdot \mathbb{P}(Y = +1 \mid X) \\ &= 2\eta(X) - 1. \end{aligned}$$

Also,

$$\begin{aligned} \mathcal{R}_{\Phi^{\mathsf{square}}}[g] - \mathcal{R}_{\Phi^{\mathsf{square}}}[g^\star] &= \mathop{\mathbb{E}}_{(X,Y)\sim P}[(g(X) - Y)^2 - (g^\star(X) - Y)^2] \\ &= \mathop{\mathbb{E}}_{(X,Y)\sim P}[g(X)^2 - 2g(X)Y - g^\star(X)^2 + 2g^\star(X)Y] \\ &= \mathop{\mathbb{E}}_{X}\left[\mathop{\mathbb{E}}_{Y}[g(X)^2 - 2g(X)Y - g^\star(X)^2 + 2g^\star(X)Y \mid X]\right] \\ &= \mathop{\mathbb{E}}_{X}\left[g(X)^2 - 2g(X)\mathop{\mathbb{E}}_{Y}[Y \mid X] - g^\star(X)^2 + 2g^\star(X)\mathop{\mathbb{E}}_{Y}[Y \mid X]\right] \\ &= \mathop{\mathbb{E}}_{X}\left[g(X)^2 - 2g(X) + g^\star(X)^2\right] \\ &= \mathop{\mathbb{E}}_{X}\left[(g(X) - g^\star(X))^2\right]. \end{aligned}$$

Minimize surrogate loss $\overset{?}{\Rightarrow}$ Minimize original loss

However, we should not forget that we have changed the optimization problem from minimizing $\mathcal{R}_{\Phi^{0\text{-}1}}$ to $\mathcal{R}_{\Phi}$.

Is this valid? Does the following implication hold?

$$\mathcal{R}_{\Phi}[g] - \mathcal{R}^{\star}_{\Phi} = 0 \quad \overset{?}{\Rightarrow} \quad \mathcal{R}_{\Phi^{0\text{-}1}}[g] - \mathcal{R}^{\star}_{\Phi^{0\text{-}1}} = 0$$

In general, no.

Since $\Phi^{0\text{-}1} \leq \gamma\Phi$ for some $\gamma > 0$, if $\mathcal{R}^{\star}_{\Phi} = 0$, then $\mathcal{R}^{\star}_{\Phi^{0\text{-}1}} = 0$ and

$$\mathcal{R}_{\Phi}[g] = 0 \quad \Rightarrow \quad \mathcal{R}_{\Phi^{0\text{-}1}}[g] = 0.$$

However, if $\mathcal{R}^{\star}_{\Phi} > 0$, the desired implication does not hold in general.

When is minimizing $\mathcal{R}_\Phi$ valid?

We shall now study conditions that ensure:

$$\operatorname*{argmin}_g \mathcal{R}_\Phi[g] \subseteq \operatorname*{argmin}_g \mathcal{R}_{\Phi^{\text{0-1}}}[g].$$

If so, then (exactly) minimizing $\mathcal{R}_\Phi$ provides a minimizer to $\mathcal{R}_{\Phi^{\text{0-1}}}$, the actual risk that we care about, i.e.,

$$\mathcal{R}_\Phi[g] - \mathcal{R}_\Phi^\star = 0 \quad \Rightarrow \quad \mathcal{R}_{\Phi^{\text{0-1}}}[g] - \mathcal{R}_{\Phi^{\text{0-1}}}^\star = 0.$$

Conditional Φ-risk

For any $g \colon \mathcal{X} \to \mathbb{R}$, define the *conditional* Φ*-risk* as

$$\begin{aligned} \mathcal{R}_\Phi[g \,|\, X] &= \mathop{\mathbb{E}}_{Y \sim P_{Y \,|\, X}} [\Phi(Y g(X)) \,|\, X] \\ &= \eta(X)\Phi(g(X)) + (1 - \eta(X))\Phi(-g(X)). \end{aligned}$$

(Of course, $\mathbb{E}_X[\mathcal{R}_\Phi[g \,|\, X]] = \mathcal{R}_\Phi[g]$.)

Let

$$C_\Phi(\alpha; \eta) = \eta\Phi(\alpha) + (1 - \eta)\Phi(-\alpha).$$

Then,

$$\mathcal{R}_\Phi[g \,|\, X] = C_\Phi(g(X); \eta(X)).$$

Bayes predictor from conditional Φ-risk

Recall that the Bayes predictor was obtained by

$$g_\Phi^\star(X) \in \operatorname*{argmin}_{\alpha\in\mathbb{R}} \mathop{\mathbb{E}}_{Y\sim P_{Y\mid X}}[\Phi(Y\alpha)\mid X] = \operatorname*{argmin}_{\alpha\in\mathbb{R}} C_\Phi(\alpha;\eta(X)).$$

For the true 0-1 loss, we have

$$\begin{aligned}\operatorname*{argmin}_{\alpha\in\mathbb{R}} C_{\Phi^{0\text{-}1}}(\alpha;\eta(X)) &= \operatorname*{argmin}_{\alpha\in\mathbb{R}} \left\{\eta(X)\mathbf{1}_{\{\alpha\le 0\}} + (1-\eta(X))\mathbf{1}_{\{\alpha\ge 0\}}\right\} \\ &= \begin{cases} \alpha > 0 & \text{if } \eta(X) > 1/2 \\ \alpha < 0 & \text{if } \eta(X) < 1/2. \end{cases}\end{aligned}$$

(For simplicity, assume $\eta(X) \neq 1/2$ with probability 1.) I.e., it is optimal to output $\alpha > 0$ if $Y = +1$ is more likely and $\alpha < 0$ if $Y = -1$ is more likely. Does this hold for the surrogate loss function?

Calibrated surrogate loss

We say a surrogate loss Φ is *classification calibrated* or *calibrated* if

$$\operatorname*{argmin}_{\alpha\in\mathbb{R}} C_\Phi(\alpha;\eta(X)) \subseteq \operatorname*{argmin}_{\alpha\in\mathbb{R}} C_{\Phi^{\text{0-1}}}(\alpha;\eta(X)) = \begin{cases} \alpha > 0 & \text{if } \eta(X) > 1/2 \\ \alpha < 0 & \text{if } \eta(X) < 1/2. \end{cases}$$

Lemma

Let Φ be classification calibrated. Then,

$$\operatorname*{argmin}_{g} \mathcal{R}_\Phi[g] \subseteq \operatorname*{argmin}_{g} \mathcal{R}_{\Phi^{\text{0-1}}}[g].$$

Proof. Let $g^\star_\Phi \in \operatorname{argmin}_g \mathcal{R}_\Phi[g]$. Then,

$$g^\star_\Phi(X) \in \operatorname*{argmin}_{\alpha\in\mathbb{R}} C_\Phi(\alpha;\eta(X))$$

for P-almost all X. Then,

$$g^\star_\Phi(X) \in \operatorname*{argmin}_{\alpha\in\mathbb{R}} C_{\Phi^{\text{0-1}}}(\alpha;\eta(X))$$

for P-almost all X, and we conclude

$$g^\star_\Phi \in \operatorname*{argmin}_{g} \mathcal{R}_{\Phi^{\text{0-1}}}[g].$$

□

Bayes predictor for square loss is optimal for 0-1 loss

Recall that

$$g^\star_{\Phi^{\text{square}}}(X) = 2\eta(X) - 1.$$

Since $g^\star_{\Phi^{\text{square}}}(X) > 0$ if $\eta(X) > 1/2$ and vice versa,

$$g^\star_{\Phi^{\text{square}}}(X) \in \operatorname*{argmin}_{\alpha \in \mathbb{R}} C_{\Phi^{\text{0-1}}}(\alpha; \eta(X)).$$

Therefore,

$$g^\star_{\Phi^{\text{square}}} \in \operatorname*{argmin}_{g} \mathcal{R}_{\Phi^{\text{0-1}}}[g].$$

How about

$$g^\star_{\Phi^{\text{logistic}}} \overset{?}{\in} \operatorname*{argmin}_{g} \mathcal{R}_{\Phi^{\text{0-1}}}, \qquad g^\star_{\Phi^{\text{hinge}}} \overset{?}{\in} \operatorname*{argmin}_{g} \mathcal{R}_{\Phi^{\text{0-1}}}$$

Calibrated surrogate loss

Theorem

Let $\Phi\colon \mathbb{R} \to \mathbb{R}$ be convex. If Φ is differentiable at 0 and $\Phi'(0) < 0$, then Φ is classification-calibrated.

Proof. Convexity of Φ implies $C_\Phi(\alpha; \eta)$ is convex in α for any fixed $\eta \in [0, 1]$. If $\eta > 1/2$, then

$$\frac{d}{d\alpha} C_\Phi(\alpha; \eta)\big|_{\alpha=0} = \eta\Phi'(0) - (1-\eta)\Phi'(0) < 0.$$

Therefore, $\operatorname{argmin}_{\alpha \in \mathbb{R}} C_\Phi(\alpha; \eta) \subseteq (0, \infty)$ by convexity.

If $\eta < 1/2$, then

$$\frac{d}{d\alpha} C_\Phi(\alpha; \eta)\big|_{\alpha=0} = \eta\Phi'(0) - (1-\eta)\Phi'(0) > 0.$$

Therefore, $\operatorname{argmin}_{\alpha \in \mathbb{R}} C_\Phi(\alpha; \eta) \subseteq (-\infty, 0)$ by convexity. $\square$

Calibrated surrogate loss

Therefore, all three surrogate losses are calibrated, and

$$
\begin{aligned}
g^\star_{\Phi^{\text{logistic}}} &\in \operatorname*{argmin}_g \mathcal{R}_{\Phi^{\text{0-1}}}[g] \\
g^\star_{\Phi^{\text{hinge}}} &\in \operatorname*{argmin}_g \mathcal{R}_{\Phi^{\text{0-1}}}[g] \\
g^\star_{\Phi^{\text{square}}} &\in \operatorname*{argmin}_g \mathcal{R}_{\Phi^{\text{0-1}}}[g].
\end{aligned}
$$

Outline

Calibration

$$\mathcal{R}_{\Phi^{0\text{-}1}} - \mathcal{R}^\star_{\Phi^{0\text{-}1}} \overset{?}{\leq} H(\mathcal{R}_\Phi - \mathcal{R}^\star_\Phi)$$

Calibration function with convex losses

Ridge least squares regression

ℓ^2-regularized estimation

When is <u>approximately</u> minimizing $\mathcal{R}_\Phi$ valid?

If Φ is calibrated, then

$$\mathcal{R}_\Phi[g] - \mathcal{R}^\star_\Phi = 0 \quad \Rightarrow \quad \mathcal{R}_{\Phi^{0\text{-}1}}[g] - \mathcal{R}^\star_{\Phi^{0\text{-}1}} = 0.$$

However, do we have?

$$\mathcal{R}_\Phi[g] - \mathcal{R}^\star_\Phi < \text{small} \quad \Rightarrow \quad \mathcal{R}_{\Phi^{0\text{-}1}}[g] - \mathcal{R}^\star_{\Phi^{0\text{-}1}} < \text{small}$$

After all, we can only hope to approximately minimize $\mathcal{R}_\Phi$.

$\mathcal{R}_{\Phi^{0\text{-}1}} - \mathcal{R}^\star_{\Phi^{0\text{-}1}} \overset{?}{\leq} H(\mathcal{R}_\Phi - \mathcal{R}^\star_\Phi)$

$$\mathcal{R}_{\Phi^{\text{0-1}}}[g] - \mathcal{R}^\star_{\Phi^{\text{0-1}}} \le \mathcal{R}_{\Phi^{\textbf{hinge}}}[g] - \mathcal{R}^\star_{\Phi^{\textbf{hinge}}}$$

For the hinge loss, we can carry out the analysis with direct arguments.

Recall,

$$C_{\Phi^{\text{0-1}}}(\alpha;\eta) = \eta \mathbf{1}_{\{\alpha \le 0\}} + (1-\eta)\mathbf{1}_{\{\alpha \ge 0\}}$$
$$C_{\Phi^{\text{hinge}}}(\alpha;\eta) = \eta(1-\alpha)_+ + (1-\eta)(1+\alpha)_+.$$

With direct calculations, we get

$$\inf_{\alpha\in\mathbb{R}} C_{\Phi^{\text{0-1}}}(\alpha;\eta) = \min\{\eta, 1-\eta\}, \qquad \inf_{\alpha\in\mathbb{R}} C_{\Phi^{\text{hinge}}}(\alpha;\eta) = 2\min\{\eta, 1-\eta\}.$$

With direct (albeit tedious) arguments, we can show

$$C_{\Phi^{\text{0-1}}}(\alpha;\eta) - \inf_{\alpha\in\mathbb{R}} C_{\Phi^{\text{0-1}}}(\alpha;\eta) \le C_{\Phi^{\text{hinge}}}(\alpha;\eta) - \inf_{\alpha\in\mathbb{R}} C_{\Phi^{\text{hinge}}}(\alpha;\eta)$$

for all $\alpha \in \mathbb{R}$ and $\eta \in [0,1]$, which implies

$$\mathcal{R}_{\Phi^{\text{0-1}}}[g] - \mathcal{R}^\star_{\Phi^{\text{0-1}}} \le \mathcal{R}_{\Phi^{\text{hinge}}}[g] - \mathcal{R}^\star_{\Phi^{\text{hinge}}}.$$

$$\mathcal{R}_{\Phi^{0\text{-}1}}[g] - \mathcal{R}^\star_{\Phi^{0\text{-}1}} \not\lesssim \mathcal{R}_{\Phi^{\text{logistic}}}[g] - \mathcal{R}^\star_{\Phi^{\text{logistic}}}$$

For the logistic loss, we have

$$C_{\Phi^{0\text{-}1}}(\alpha;\eta) \le \frac{1}{\log 2} C_{\Phi^{\text{logistic}}}(\alpha;\eta)$$

However,

$$C_{\Phi^{0\text{-}1}}(\alpha;\eta) - \inf_{\alpha\in\mathbb{R}} C_{\Phi^{0\text{-}1}}(\alpha;\eta) \not\le \gamma\Big(C_{\Phi^{\text{logistic}}}(\alpha;\eta) - \inf_{\alpha\in\mathbb{R}} C_{\Phi^{\text{logistic}}}(\alpha;\eta)\Big)$$

for any constant $\gamma > 0$, and we cannot proceed with the same argument.

The same problem arises with the square loss.

Lemma

Let $g^\star \in \operatorname{argmin}_g \mathcal{R}_{\Phi^{0\text{-}1}}[g]$. *Then,*

$$\begin{aligned}\mathcal{R}_{\Phi^{0\text{-}1}}[g] - \mathcal{R}_{\Phi^{0\text{-}1}}[g^\star] &= \mathbb{E}\big[\mathbf{1}_{\{g(X)g^\star(X)<0\}}|2\eta(X)-1|\big] \\ &\le \mathbb{E}\big[\mathbf{1}_{\{g(X)g^\star(X)<0\}}|2\eta(X)-1-b(g(X))|\big]\end{aligned}$$

for any $b\colon \mathbb{R}\to\mathbb{R}$ *such that* $\operatorname{sign}(x)\operatorname{sign}(b(x)) \ge 0$ *for all* $x \in \mathbb{R}$.

Proof. The first claim follows from

$$
\begin{aligned}
&\mathcal{R}_{\Phi^{0\text{-}1}}[g] - \mathcal{R}_{\Phi^{0\text{-}1}}[g^\star] \\
&\quad = \mathbb{E}\Big[\mathbb{E}\big[\mathbf{1}_{\{\mathrm{sign}(g(X))\neq Y\}} - \mathbf{1}_{\{\mathrm{sign}(g^\star(X))\neq Y\}} \,\big|\, X\big]\Big] \\
&\quad = \mathbb{E}\Big[\mathbb{E}\big[- \mathbf{1}_{\{g(X)>0,\, g^\star(X)<0\}}\mathbf{1}_{\{Y=+1\}} + \mathbf{1}_{\{g(X)>0,\, g^\star(X)<0\}}\mathbf{1}_{\{Y=-1\}} \\
&\qquad\qquad + \mathbf{1}_{\{g(X)<0,\, g^\star(X)>0\}}\mathbf{1}_{\{Y=+1\}} - \mathbf{1}_{\{g(X)<0,\, g^\star(X)>0\}}\mathbf{1}_{\{Y=-1\}} \,\big|\, X\big]\Big] \\
&\quad = \mathbb{E}\Big[- \mathbf{1}_{\{g(X)>0,\, g^\star(X)<0\}}\eta(X) + \mathbf{1}_{\{g(X)>0,\, g^\star(X)<0\}}(1-\eta(X)) \\
&\qquad\qquad + \mathbf{1}_{\{g(X)<0,\, g^\star(X)>0\}}\eta(X) - \mathbf{1}_{\{g(X)<0,\, g^\star(X)>0\}}(1-\eta(X))\Big] \\
&\quad = \mathbb{E}\Big[\mathbf{1}_{\{g(X)>0,\, g^\star(X)<0\}}(1-2\eta(X)) - \mathbf{1}_{\{g(X)<0,\, g^\star(X)>0\}}(1-2\eta(X))\Big] \\
&\quad = \mathbb{E}\Big[\mathbf{1}_{\{g(X)>0,\, g^\star(X)<0\}}|1-2\eta(X)| + \mathbf{1}_{\{g(X)<0,\, g^\star(X)>0\}}|1-2\eta(X)|\Big] \\
&\quad = \mathbb{E}\Big[\mathbf{1}_{\{g(X)g^\star(X)<0\}}|1-2\eta(X)|\Big],
\end{aligned}
$$

where we use the fact that $g^\star(X) < 0$ implies $\eta(X) < 1/2$.

For the second claim,

$$
\begin{aligned}
&\mathbb{E}\big[\mathbf{1}_{\{g(X)g^\star(X)<0\}}|2\eta(X)-1|\big] \\
&\quad = \mathbb{E}\big[\mathbf{1}_{\{g(X)g^\star(X)<0,\, g^\star(X)>0,\, \eta(X)>1/2\}}(2\eta(X)-1)\big] \\
&\qquad\quad + \mathbb{E}\big[\mathbf{1}_{\{g(X)g^\star(X)<0,\, g^\star(X)<0,\, \eta(X)<1/2\}}(-2\eta(X)+1)\big] \\
&\quad \le \mathbb{E}\big[\mathbf{1}_{\{g(X)g^\star(X)<0,\, g^\star(X)>0,\, \eta(X)>1/2\}}(2\eta(X)-1-b(g(X)))\big] \\
&\qquad\quad + \mathbb{E}\big[\mathbf{1}_{\{g(X)g^\star(X)<0,\, g^\star(X)<0,\, \eta(X)<1/2\}}(-2\eta(X)+1+b(g(X)))\big] \\
&\quad = \mathbb{E}\big[\mathbf{1}_{\{g(X)g^\star(X)<0,\, g^\star(X)>0,\, \eta(X)>1/2\}}|2\eta(X)-1-b(g(X))|\big] \\
&\qquad\quad + \mathbb{E}\big[\mathbf{1}_{\{g(X)g^\star(X)<0,\, g^\star(X)<0,\, \eta(X)<1/2\}}|2\eta(X)-1-b(g(X))|\big] \\
&\quad = \mathbb{E}\big[\mathbf{1}_{\{g(X)g^\star(X)<0\}}|2\eta(X)-1-b(g(X))|\big].
\end{aligned}
$$

$\square$

Square loss

Equipped with this lemma, we can now analyze the relationship between $\mathcal{R}_{\Phi^{0\text{-}1}}[g] - \mathcal{R}^\star_{\Phi^{0\text{-}1}}$ and $\mathcal{R}_{\Phi^{\text{square}}}[g] - \mathcal{R}^\star_{\Phi^{\text{square}}}$:

$$\begin{aligned}
\mathcal{R}_{\Phi^{0\text{-}1}}[g] - \mathcal{R}_{\Phi^{0\text{-}1}}[g^\star] &\leq \mathbb{E}\big[\mathbf{1}_{\{g(X)g^\star(X)<0\}}|2\eta(X) - 1 - g(X)|\big] \\
&\leq \Big(\mathbb{E}\big[\mathbf{1}_{\{g(X)g^\star(X)<0\}}|\underbrace{2\eta(X) - 1}_{=g^\star(X)} - g(X)|^2\big]\Big)^{1/2} \\
&\leq \big(\mathbb{E}[|g^\star(X) - g(X)|^2]\big)^{1/2} \\
&= \Big(\mathcal{R}_{\Phi^{\text{square}}}[g] - \mathcal{R}_{\Phi^{\text{square}}}[g^\star]\Big)^{1/2},
\end{aligned}$$

where the second inequality follows from Jensen.

Therefore,

$$\mathcal{R}_{\Phi^{\text{square}}} - \mathcal{R}^\star_{\Phi^{\text{square}}} < \text{small} \quad \Rightarrow \quad \mathcal{R}_{\Phi^{0\text{-}1}} - \mathcal{R}^\star_{\Phi^{0\text{-}1}} < \sqrt{\text{small}}.$$

Logistic loss

Lemma

For any $x, u \in \mathbb{R}$

$$\log(e^{-x/2} + e^{x/2}) - ux - \inf_{x \in \mathbb{R}} \{\log(e^{-x/2} + e^{x/2}) - ux\} \geq 2\Big(u - \frac{e^x - 1}{2(e^x + 1)}\Big)^2.$$

Proof. A brute-force proof:

$$\begin{aligned}\inf_x \{\log(e^{-x/2} + e^{x/2}) - ux\} &= \begin{cases} \frac{1}{2}(1-2u)\log\frac{1+2u}{1-2u} + \log\frac{2}{1+2u} & \text{if } -2 < u < 2 \\ -\infty & \text{otherwise.} \end{cases} \\ &\leq \log(e^{-x/2} + e^{x/2}) - ux - 2\Big(u - \frac{e^x - 1}{2(e^x + 1)}\Big)^2 \end{aligned}$$

with a Taylor expansion argument. Rearrange the inequality to conclude the statement. (Better proof later.) □

Logistic loss

Recall that

$$\Phi^{\text{logistic}}(u) = \log(1 + e^{-u}).$$

Then,

$$\begin{aligned} C_{\Phi^{\text{logistic}}}(\alpha; \eta) &= \eta \log(1 + e^{-\alpha}) + (1 - \eta) \log(1 + e^{\alpha}) \\ &= \log(e^{-\alpha/2} + e^{\alpha/2}) - \frac{2\eta - 1}{2}\alpha \end{aligned}$$

Appealing to the previous lemma, we have

$$C_{\Phi^{\text{logistic}}}(\alpha; \eta) - \inf_{\alpha \in \mathbb{R}} C_{\Phi^{\text{logistic}}}(\alpha; \eta) \geq \frac{1}{2}\Big(2\eta - 1 - \frac{e^{\alpha} - 1}{e^{\alpha} + 1}\Big)^2.$$

Logistic loss

Plug in $\alpha \leftarrow g(X)$ and $\eta \leftarrow \eta(X)$, and take the expectation to get

$$\begin{aligned}
\mathcal{R}_{\Phi^{\text{logistic}}}[g] - \mathcal{R}^\star_{\Phi^{\text{logistic}}} &\geq \frac{1}{2}\mathbb{E}\Big[\Big(2\eta(X) - 1 - \frac{e^{g(X)}-1}{e^{g(X)}+1}\Big)^2\Big] \\
&\geq \frac{1}{2}\Big(\mathbb{E}\Big[\Big|2\eta(X) - 1 - \frac{e^{g(X)}-1}{e^{g(X)}+1}\Big|\Big]\Big)^2 \\
&\geq \frac{1}{2}\Big(\mathcal{R}_{\Phi^{\text{0-1}}}[g] - \mathcal{R}^\star_{\Phi^{\text{0-1}}}\Big)^2.
\end{aligned}$$

Therefore, we conclude

$$\mathcal{R}_{\Phi^{\text{0-1}}}[g] - \mathcal{R}^\star_{\Phi^{\text{0-1}}} \leq \sqrt{2}\Big(\mathcal{R}_{\Phi^{\text{logistic}}}[g] - \mathcal{R}^\star_{\Phi^{\text{logistic}}}\Big)^{1/2}$$

the same (up to constant) guarantee as for the square loss.

Calibration function

We established guarantees of the form

$$\mathcal{R}_{\Phi^{0\text{-}1}} - \mathcal{R}^\star_{\Phi^{0\text{-}1}} \overset{?}{\leq} H(\mathcal{R}_\Phi - \mathcal{R}^\star_\Phi),$$

where H is a monotonically increasing function satisfying $H(0) = 0$. H is called the *calibration function*.

The guarantee for the hinge loss is better than the guarantee for the square or logistic loss. However, we will later see that the hinge loss is harder to optimize due to its non-differentiability. So there is a trade-off.

Impact on approximation errors

So far, our analysis has been carried out without any restriction on the set of functions.

In practice, however, we use a restricted function class $\mathcal{F}$ (often with a controlled Rademacher complexity). The choice of the surrogate loss Φ affects the Bayes predictor (even though the set of Bayes predictor for $\Phi^{0\text{-}1}$ is always the same), so the approximation error is affected by the choice of Φ.

In particular,

$$\begin{aligned}
g^\star_{\Phi^{\text{hinge}}}(X) &= \operatorname{sign}(2\eta(X) - 1) \\
g^\star_{\Phi^{\text{logistic}}}(X) &= \operatorname{atanh}(2\eta(X) - 1) \\
g^\star_{\Phi^{\text{square}}}(X) &= 2\eta(X) - 1.
\end{aligned}$$

If Φ admits a $g^\star_\Phi$ that is well approximated by $\mathcal{F}$, that is a reason to favor Φ. (Having a favorable calibration function and the ease of optimization are two other reasons to favor a choice of Φ.)

Outline

Calibration

$\mathcal{R}_{\Phi^{0\text{-}1}} - \mathcal{R}^{\star}_{\Phi^{0\text{-}1}} \overset{?}{\leq} H(\mathcal{R}_{\Phi} - \mathcal{R}^{\star}_{\Phi})$

Calibration function with convex losses

Ridge least squares regression

ℓ^2-regularized estimation

Logistic loss

Lemma

For any $x, u \in \mathbb{R}$

$$\log(e^{-x/2} + e^{x/2}) - ux - \inf_{x \in \mathbb{R}} \{\log(e^{-x/2} + e^{x/2}) - ux\} \geq 2(u - b(x))^2,$$

where $b \colon \mathbb{R} \to \mathbb{R}$ *is a sign-preserving function.*

Better Proof. Note that

$$f(x) = \log(e^{-x/2} + e^{x/2})$$

is a convex L-smooth function with $L = 1/4$. (Easy to check that $0 \leq f''(x) \leq 1/4$ for all $x \in \mathbb{R}$.) Then, by the Fenchel–Young inequality for smooth convex functions, we have

$$f(x) + f^*(u) - ux \geq \frac{1}{2L}(u - f'(x))^2.$$

Finally, it is straightforward to verify $f'(0) = 0$ and f' is strictly increasing. □

Calibration functions for square and logistic losses

Assume

$$\Phi(u) = a(u) - \gamma u + \beta,$$

where $a(0) = 0$, a is convex L-smooth, a is even, $\gamma > 0$, and $\beta \in \mathbb{R}$.
Recall

$$\Phi^{\text{square}}(u) = (1-u)^2 = u^2 - 2u + 1 \qquad \text{(2-smooth)}$$

$$\Phi^{\text{logistic}}(u) = \log(1 + e^{-u}) = \log(e^{-u/2} + e^{u/2}) - \frac{1}{2}u \qquad \text{($\tfrac{1}{4}$-smooth)}$$

Then,

$$\begin{aligned} C_\Phi(\alpha; \eta) &= \eta\Phi(\alpha) + (1-\eta)\Phi(-\alpha) + \beta \\ &= a(\alpha) - \gamma(2\eta - 1)\alpha \end{aligned}$$

Using Fenchel–Young for smooth convex functions, we have

$$C_\Phi(\alpha; \eta) - \inf_{\alpha \in \mathbb{R}} C_\Phi(\alpha; \eta) \geq \frac{\gamma^2}{2L}\Big(2\eta - 1 - \tfrac{1}{\gamma}a'(\alpha)\Big)^2.$$

Calibration functions for square and logistic losses

Plug in $\alpha \leftarrow g(X)$ and $\eta \leftarrow \eta(X)$, and take the expectation to get

$$\begin{aligned}
\mathcal{R}_\Phi[g] - \mathcal{R}_\Phi^\star &\geq \frac{\gamma^2}{2L}\mathbb{E}\Big[\Big(2\eta(X) - 1 - \tfrac{1}{\gamma}a'(g(X))\Big)^2\Big] \\
&\geq \frac{\gamma^2}{2L}\Big(\mathbb{E}\Big[\Big|2\eta(X) - 1 - \tfrac{1}{\gamma}a'(g(X))\Big|\Big]\Big)^2 \\
&\geq \frac{\gamma^2}{2L}\Big(\mathcal{R}_{\Phi^{0\text{-}1}}[g] - \mathcal{R}_{\Phi^{0\text{-}1}}^\star\Big)^2.
\end{aligned}$$

Therefore, we conclude

$$\mathcal{R}_{\Phi^{0\text{-}1}}[g] - \mathcal{R}_{\Phi^{0\text{-}1}}^\star \leq \frac{\sqrt{2L}}{\gamma}\Big(\mathcal{R}_\Phi[g] - \mathcal{R}_\Phi^\star\Big)^{1/2}.$$

Calibration function with conjugate functions

Consider surrogate loss function $\Phi\colon \mathbb{R} \to \mathbb{R}$ of the form

$$\Phi(\alpha) = a(\alpha) - \alpha,$$

where $a(0) = 0$, a is even, and a is convex. Then,

$$a^*\left(\mathcal{R}_{\Phi^{0\text{-}1}}[g] - \mathcal{R}^\star_{\Phi^{0\text{-}1}}\right) \leq \mathcal{R}_\Phi[g] - \mathcal{R}^*_\Phi$$

for any $g\colon \mathcal{X} \to \mathbb{R}$, where a^* is the conjugate function of a.

First, note that $a \geq 0$ since a is convex, even, and $a(0) = 0$. Also,

$$a^*(0) = -\inf_{\alpha \in \mathbb{R}} \{a(\alpha)\} = -a(0) = 0.$$

Calibration function with conjugate functions

Recall

$$\begin{aligned}\mathcal{R}_{\Phi^{0\text{-}1}}[g] - \mathcal{R}^\star_{\Phi^{0\text{-}1}} &= \mathbb{E}\big[\mathbf{1}_{\{g(X)g^\star(X)<0\}}|2\eta(X)-1|\big] \\ &= \mathbb{E}\left[\mathbf{1}_{\{g(X)(2\eta(X)-1)<0\}}|2\eta(X)-1|\right].\end{aligned}$$

By Jensen's inequality,

$$\begin{aligned}a^*\big(\mathcal{R}_{\Phi^{0\text{-}1}}[g] - \mathcal{R}^\star_{\Phi^{0\text{-}1}}\big) &= a^*\Big(\mathbb{E}\big[\mathbf{1}_{\{g(X)(2\eta(X)-1)<0\}}|2\eta(X)-1|\big]\Big) \\ &\leq \mathbb{E}\Big[a^*\big(\mathbf{1}_{\{g(X)(2\eta(X)-1)<0\}}|2\eta(X)-1|\big)\Big] \\ &= \mathbb{E}\Big[\mathbf{1}_{\{g(X)(2\eta(X)-1)<0\}}a^*(|2\eta(X)-1|)\Big],\end{aligned}$$

where the final equality uses $a^*(0) = 0$.

Calibration function with conjugate functions

Recall our definition

$$C_\Phi(\alpha;\eta) = \eta\Phi(\alpha) + (1-\eta)\Phi(-\alpha) = a(\alpha) - (2\eta-1)\alpha.$$

Then,

$$\begin{aligned}
a^*\left(|2\eta(X)-1|\right) &= -\inf_{\alpha\in\mathbb{R}} \{a(\alpha) - |2\eta(X)-1|\alpha\}) \\
&= -\inf_{\alpha\in\mathbb{R}} \{a(\alpha) - (2\eta(X)-1)\alpha\} \quad (\because a \text{ is even}) \\
&= \inf_{(2\eta(X)-1)\alpha\le 0} \{a(\alpha) - (2\eta(X)-1)\alpha\} - \inf_{\alpha\in\mathbb{R}} \{a(\alpha) - (2\eta(X)-1)\alpha\} \\
&= \inf_{(2\eta(X)-1)\alpha\le 0} \{a(\alpha) - (2\eta(X)-1)\alpha\} - \inf_{\alpha\in\mathbb{R}} C_\Phi(\alpha;\eta(X)),
\end{aligned}$$

where the third equality follows from $a \ge 0$.

To clarify, $\inf_{(2\eta(X)-1)\alpha\le 0}$ takes inf over $\{\alpha\in\mathbb{R}\,|\,(2\eta(X)-1)\alpha\le 0\}$.

Calibration function with conjugate functions

To save space, write

$$\mathbf{1}_{\{\cdots\}} = \mathbf{1}_{\{g(X)(2\eta(X)-1)<0\}}$$

within this page. Combining the bounds, we get

$$\begin{aligned}
&a^*(\mathcal{R}_{\Phi^{0\text{-}1}}[g] - \mathcal{R}^\star_{\Phi^{0\text{-}1}}) \\
&\le \mathbb{E}\left[\mathbf{1}_{\{\cdots\}}\left(\inf_{(2\eta(X)-1)\alpha\le 0}\{a(\alpha) - (2\eta(X)-1)\alpha\} - \inf_{\alpha\in\mathbb{R}} C_\Phi(\alpha;\eta(X))\right)\right] \\
&\quad \le \mathbb{E}\left[\mathbf{1}_{\{\cdots\}}\left(\big(a(g(X)) - (2\eta(X)-1)g(X)\big) - \inf_{\alpha\in\mathbb{R}} C_\Phi(\alpha;\eta(X))\right)\right] \\
&\quad = \mathbb{E}\left[\mathbf{1}_{\{\cdots\}}\left(C_\Phi(g(X),\eta(X)) - \inf_{\alpha\in\mathbb{R}} C_\Phi(\alpha;\eta(X))\right)\right] \\
&\quad \le \mathbb{E}\left[C_\Phi(g(X),\eta(X)) - \inf_{\alpha\in\mathbb{R}} C_\Phi(\alpha;\eta(X))\right] \\
&\quad = \mathcal{R}_\Phi[g] - \mathcal{R}^*_\Phi.
\end{aligned}$$

Outline

Calibration

$\mathcal{R}_{\Phi^{0\text{-}1}} - \mathcal{R}^\star_{\Phi^{0\text{-}1}} \overset{?}{\leq} H(\mathcal{R}_\Phi - \mathcal{R}^\star_\Phi)$

Calibration function with convex losses

Ridge least squares regression

ℓ^2-regularized estimation

Ridge regression

Consider the *ridge regression* problem with $\mu > 0$

$$\underset{\theta \in \mathbb{R}^d}{\text{minimize}} \quad \tfrac{1}{N}\|Y - \Phi\theta\|^2 + \mu\|\theta\|_2^2,$$

which has the minimizer

$$\hat{\theta}_\mu = \frac{1}{N}(\widehat{\Sigma} + \mu I)^{-1}\Phi^\intercal Y = (\Phi^\intercal \Phi + N\mu I)^{-1}\Phi^\intercal Y = \Phi^\intercal(\Phi\Phi^\intercal + N\mu I)^{-1} Y.$$

(Proof involving matrix inversion lemma in homework.)

Recall

$$\widehat{\Sigma} = \frac{1}{N}\Phi^\intercal \Phi \in \mathbb{R}^{d \times d}.$$

Notably, we will no longer assume that $\widehat{\Sigma}$ is invertible. Not assuming invertibility will be important when we consider kernel methods, where $d = \infty$ and $N < \infty$.

Ridge regression

Even though we consider a regularized optimization problem to obtain $\hat{\theta}_\mu$, we still consider the same (unregularized) risk

$$\mathcal{R}(\theta) = \mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_N}[\tfrac{1}{N}\|\Phi\theta - Y\|^2].$$

Theorem

For the fixed design setting, with $\hat{\theta}_\mu = \frac{1}{N}(\widehat{\Sigma} + \mu I)^{-1}\Phi^\intercal Y$ *has expected excess risk*

$$\mathbb{E}[\mathcal{R}(\hat{\theta})] - \mathcal{R}^\star = \underbrace{\mu^2\theta_\star^\intercal(\widehat{\Sigma} + \mu I)^{-2}\widehat{\Sigma}\theta_\star}_{\textit{bias}} + \underbrace{\frac{\sigma^2}{N}\mathrm{Tr}(\widehat{\Sigma}^2(\widehat{\Sigma} + \mu I)^{-2})}_{\textit{variance}}.$$

Proof. Recall that we had shown

$$\mathbb{E}[\mathcal{R}(\hat{\theta})] - \mathcal{R}_\star = \underbrace{\|\mathbb{E}[\hat{\theta}] - \theta_\star\|_{\widehat{\Sigma}}^2}_{\text{bias}} + \underbrace{\mathbb{E}[\|\hat{\theta} - \mathbb{E}[\hat{\theta}]\|_{\widehat{\Sigma}}^2]}_{\text{variance}},$$

First, we have

$$\begin{aligned}
\mathbb{E}[\hat{\theta}_\mu] &= \frac{1}{N}\mathbb{E}[(\widehat{\Sigma} + \mu I)^{-1}\Phi^\intercal(\Phi\theta_\star + \varepsilon)] \\
&= (\widehat{\Sigma} + \mu I)^{-1}\widehat{\Sigma}\theta_\star = (\widehat{\Sigma} + \mu I)^{-1}(\widehat{\Sigma} + \mu I - \mu I)\theta_\star \\
&= \theta_\star - \mu(\widehat{\Sigma} + \mu I)^{-1}\theta_\star.
\end{aligned}$$

So

$$\begin{aligned}
\text{bias} &= \|\mu(\widehat{\Sigma} + \mu I)^{-1}\theta_\star\|_{\widehat{\Sigma}}^2 = \mu^2\theta_\star^\intercal(\widehat{\Sigma} + \mu I)^{-1}\widehat{\Sigma}(\widehat{\Sigma} + \mu I)^{-1}\theta_\star \\
&= \mu^2\theta_\star^\intercal(\widehat{\Sigma} + \mu I)^{-2}\widehat{\Sigma}\theta_\star,
\end{aligned}$$

which accounts for the first term.

Next, we have

$$\hat{\theta} - \mathbb{E}[\hat{\theta}] = \frac{1}{N}(\widehat{\Sigma} + \mu I)^{-1}\Phi^{\mathsf{T}}\varepsilon.$$

So,

$$\begin{aligned}
\text{variance} &= \mathbb{E}\Big[\|\hat{\theta} - \mathbb{E}[\hat{\theta}]\|_{\widehat{\Sigma}}^2\Big] \\
&= \frac{1}{N^2}\mathbb{E}\Big[\mathrm{Tr}\big(\varepsilon^{\mathsf{T}}\Phi(\widehat{\Sigma} + \mu I)^{-1}\widehat{\Sigma}(\widehat{\Sigma} + \mu I)^{-1}\Phi^{\mathsf{T}}\varepsilon\big)\Big] \\
&= \frac{\sigma^2}{N^2}\mathrm{Tr}\big(\Phi(\widehat{\Sigma} + \mu I)^{-1}\widehat{\Sigma}(\widehat{\Sigma} + \mu I)^{-1}\Phi^{\mathsf{T}}\big) \\
&= \frac{\sigma^2}{N}\mathrm{Tr}\big((\widehat{\Sigma} + \mu I)^{-1}\widehat{\Sigma}(\widehat{\Sigma} + \mu I)^{-1}\widehat{\Sigma}\big) \\
&= \frac{\sigma^2}{N}\mathrm{Tr}\big(\widehat{\Sigma}^2(\widehat{\Sigma} + \mu I)^{-2}\big).
\end{aligned}$$

□

Should we use $\mu > 0$?

For small $\mu > 0$, we have

$$\begin{aligned}
\mathbb{E}[\mathcal{R}(\hat{\theta}_\mu)] - \mathcal{R}^\star &= \mu^2 \theta_\star^\intercal (\widehat{\Sigma} + \mu I)^{-2} \widehat{\Sigma} \theta_\star + \frac{\sigma^2}{N} \mathrm{Tr}(\widehat{\Sigma}^2 (\widehat{\Sigma} + \mu I)^{-2}) \\
&= \mathcal{O}(\mu^2) + \frac{\sigma^2}{N} \sum_{i=1}^{\min\{d,N\}} \frac{\lambda_i^2}{(\lambda_i + \mu)^2} \\
&= \mathcal{O}(\mu^2) + \frac{\sigma^2}{N} \sum_{i=1}^{\min\{d,N\}} \frac{1}{(1 + \mu/\lambda_i)^2} \\
&= \frac{\sigma^2}{N} \sum_{i=1}^{\min\{d,N\}} (1 - 2\mu\lambda_i) + \mathcal{O}(\mu^2) \\
&= \mathcal{O}(1) - \frac{2\sigma^2 \mathrm{Tr}(\widehat{\Sigma})}{N} \mu + \mathcal{O}(\mu^2).
\end{aligned}$$

So the optimal value of μ is positive.

Optimizing regularization parameter

Theorem

Assume $\theta_\star \neq 0$. *With*

$$\mu_\circ = \frac{\sigma \mathrm{Tr}(\widehat{\Sigma})^{1/2}}{\|\theta_\star\|_2 \sqrt{N}}$$

we have

$$\mathbb{E}[\mathcal{R}(\hat{\theta}_{\mu_\circ})] - \mathcal{R}^\star \leq \frac{\sigma \mathrm{Tr}(\widehat{\Sigma})^{1/2} \|\theta_\star\|_2.}{\sqrt{N}}$$

(As we will see from the proof, $\mu_\circ$ is not the exact optimum, but rather a choice that optimizes an upper bound.)

Proof. Previously, we had shown that $\mathbb{E}[\mathcal{R}(\hat{\theta}_\mu)] - \mathcal{R}^\star = \text{bias} + \text{variance}$.

First, bound the bias:

$$\text{bias} = \mu^2 \theta_\star^\intercal (\widehat{\Sigma} + \mu I)^{-2} \widehat{\Sigma} \theta_\star = \mu \theta_\star^\intercal \underbrace{(\widehat{\Sigma} + \mu I)^{-2} \mu \widehat{\Sigma}}_{\preceq \frac{1}{2} I} \theta_\star \le \frac{\mu}{2} \|\theta_\star\|^2,$$

where we use the fact that

$$\frac{\mu \lambda}{(\lambda + \mu)^2} \le \frac{1}{2} \qquad \forall\, \mu > 0,\, \lambda > 0.$$

Next, bound the variance:

$$\text{variance} = \frac{\sigma^2}{N} \operatorname{Tr}\big(\widehat{\Sigma}^2 (\widehat{\Sigma} + \mu I)^{-2}\big) = \frac{\sigma^2}{\mu N} \operatorname{Tr}\big(\widehat{\Sigma}\, \underbrace{\mu \widehat{\Sigma} (\widehat{\Sigma} + \mu I)^{-2}}_{\preceq \frac{1}{2} I}\big) \le \frac{\sigma^2}{2 \mu N} \operatorname{Tr} \widehat{\Sigma}.$$

Finally, plugging in $\mu \leftarrow \mu_\circ$ (which minimizes the upper bounds on bias + variance), we conclude the statement. $\square$

Compared to the expected excess risk of the least squares estimator without regularization

$$\mathbb{E}[\mathcal{R}(\hat{\theta}_0)] - \mathcal{R}^\star = \frac{\sigma^2 d}{N}$$

the bound with regularization

$$\mathbb{E}[\mathcal{R}(\hat{\theta}_{\mu_\circ})] - \mathcal{R}^\star \leq \frac{\sigma \operatorname{Tr}(\widehat{\Sigma})^{1/2} \|\theta_\star\|_2}{\sqrt{N}}$$

does not have an explicit dependence on d. Such bounds are said to be *dimension-independent*.

If $\|\varphi(x)\| \leq R$ for all x, then

$$\operatorname{Tr}(\widehat{\Sigma}) = \frac{1}{N} \sum_{i=1}^{N} \|\varphi(X_i)\|_2^2 \leq R^2,$$

and the only remaining (implicit) dependence on d is in $\|\theta_\star\|_2$.

However, the $\mathcal{O}(1/\sqrt{N})$-rate is slower than the $\mathcal{O}(1/N)$-rate. This is a common tradeoff in machine learning theory: a “fast rate” with bad constants vs. “slow rate” with good constants.

Outline

Calibration

$\mathcal{R}_{\Phi^{0\text{-}1}} - \mathcal{R}^\star_{\Phi^{0\text{-}1}} \overset{?}{\leq} H(\mathcal{R}_\Phi - \mathcal{R}^\star_\Phi)$

Calibration function with convex losses

Ridge least squares regression

ℓ^2-regularized estimation

Constrained problem formulation

When learning among a function class

$$\mathcal{F} = \{f_\theta \mid \|\theta\| \le D\},$$

with $D \ge 0$, the training problem

$$\underset{f \in \mathcal{F}}{\text{minimize}} \quad \hat{\mathcal{R}}[f]$$

becomes a *constrained* optimization problem

$$\begin{array}{ll} \underset{\theta \in \mathbb{R}^p}{\text{minimize}} & \hat{\mathcal{R}}[f_\theta] \\ \text{subject to} & \|\theta\|^2 \le D^2. \end{array}$$

The constrained problem can be solved with the projected (stochastic) gradient method.

Regularized problem formulation

For various reasons, however, one often prefers the *regularized* problem formulation

$$\underset{\theta \in \mathbb{R}^p}{\text{minimize}} \quad \hat{\mathcal{R}}[f_\theta] + \lambda \|\theta\|^2,$$

where $\lambda \geq 0$.

It is often said that the constrained and regularized problems are "equivalent".

Equivalence of constrained and regularized problems

Theorem

Let $\lambda \geq 0$. *Let* $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ *and* $g\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. *Let* $x^\star$ *be a solution to*

$$\underset{x \in \mathbb{R}^d}{\textit{minimize}} \quad f(x) + \lambda g(x).$$

Set $D = g(x^\star)$. *Then,* $x^\star$ *is a solution to*

$$\begin{array}{ll} \underset{x \in \mathbb{R}^d}{\textit{minimize}} & f(x) \\ \textit{subject to} & g(x) \leq D. \end{array}$$

Proof. Assume for contradiction that $x^\star$ is not a solution for the second problem. Then, there is a $\tilde{x}$ such that $f(\tilde{x}) < f(x^\star)$ and $g(\tilde{x}) \leq D = g(x^\star)$. However, this implies

$$f(\tilde{x}) + \lambda g(\tilde{x}) < f(x^\star) + \lambda g(x^\star),$$

contradicting the premise that $x^\star$ is a solution to the first problem. □

Equivalence of constrained and regularized problems

Theorem

Let $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and $g\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be convex. Let $x^\star$ be a solution to the optimization problem

$$\begin{array}{ll} \underset{x \in \mathbb{R}^d}{\text{minimize}} & f(x) \\ \text{subject to} & g(x) \leq D. \end{array}$$

Assume the strict feasibility condition: there exists an $\tilde{x} \in \operatorname{dom} f$ such that $g(\tilde{x}) < D$. Then there is a $\lambda \geq 0$ such that $x^\star$ is a solution to

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad f(x) + \lambda g(x).$$

Proof outline. Define

$$\mathcal{A} = \{(q, p) \in \mathbb{R}^2 \,|\, g(x) \le q,\ f(x) \le p,\ x \in \mathbb{R}^d\}.$$

Then, $\mathcal{A}$ is a nonempty convex set. Note that

$$\{(g(x), f(x)) \,|\, x \in \operatorname{dom} f \cap \operatorname{dom} g\} \subseteq \mathcal{A}.$$

Since $x^\star$ is a solution, $(q^\star, p^\star) = (g(x^\star), f(x^\star)) \in \partial\mathcal{A}$, since $(q^\star, p^\star - \varepsilon) \notin \mathcal{A}$ no matter how small $\varepsilon > 0$ is.

Then, there is a supporting hyperplane: there is $(\tilde{\lambda}, \tilde{\mu}) \ne 0$ such that

$$\tilde{\lambda} q + \tilde{\mu} p \le \tilde{\lambda} q^\star + \tilde{\mu} p^\star, \qquad \forall\, (q, p) \in \mathcal{A}.$$

If $\tilde{\mu} = 0$, then set $\lambda = \tilde{\lambda}$ and $t^\star = \lambda q^\star$. If $\tilde{\mu} \ne 0$, then set $\lambda = \tilde{\lambda}/\tilde{\mu}$ and $t^\star = \lambda q^\star + p^\star$. There are four cases for the half-spaces containing $\mathcal{A}$ with $(q^\star, p^\star)$ at the boundary

- $\{(q, p) \in \mathbb{R}^2 \,|\, \lambda q \ge t^\star\}$ with $\lambda \ne 0$: Vertical half-space.
- $\{(q, p) \in \mathbb{R}^2 \,|\, \lambda q + p \le t^\star\}$: Lower half-space.
- $\{(q, p) \in \mathbb{R}^2 \,|\, \lambda q + p \ge t^\star\}$, $\lambda < 0$: Upper-left half-space.
- $\{(q, p) \in \mathbb{R}^2 \,|\, \lambda q + p \ge t^\star\}$, $\lambda \ge 0$: Upper-right half-space.

Lower half-space and upper-left half-space are not possible.

Assume

$$\mathcal{A} \subseteq \{(q,p) \in \mathbb{R}^2 \,|\, \lambda q \geq t^\star\}$$

with $\lambda \neq 0$. If $\lambda < 0$, this is a vertical-left half-space, but that is impossible. So, $\lambda > 0$ and

$$A \subseteq \{(q,p) \in \mathbb{R}^2 \,|\, q \geq t^\star/\lambda = q^\star = g(x^\star)\}.$$

This implies $g(x) \geq g(x^\star)$ for all $x \in \operatorname{dom} f$.

If $g(x^\star) = D$, the existence of $\tilde{x}$ draws a contradiction.

If $g(x^\star) < D$, then the constraint is inactive, and $x^\star$ is a solution to

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad f(x),$$

i.e., the statement holds with $\lambda = 0$.

Otherwise, consider the upper-right half-space case:

$$\mathcal{A} \subseteq \{(q,p) \in \mathbb{R}^2 \,|\, \lambda q + p \geq t^\star\}$$

with $\lambda \geq 0$.

This implies

$$\lambda g(x) + f(x) \geq t^\star = \lambda q^\star + p^\star = \lambda g(x^\star) + f(x^\star)$$

and $x^\star$ is a solution to

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad f(x) + \lambda g(x).$$

$\square$